# Save hydrological observations! Return period estimation without data decimation

Elena Volpi[a,*], Aldo Fiori[a], Salvatore Grimaldi[b,c], Federico Lombardo[d], Demetris Koutsoyiannis[e]

[a]*Department of Engineering, University of Roma Tre, Rome, Italy*
[b]*Department for Innovation in Biological, Agro-food and Forest systems (DIBAF), University of Tuscia, Viterbo, Italy*
[c]*Honors Center of Italian Universities (H2CU), Sapienza University of Rome, Rome, Italy*
[d]*Dipartimento di Ingegneria Civile, Edile e Ambientale, Sapienza Universitá di Roma, Rome, Italy*
[e]*Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Greece*

**Abstract**

The concept of return period and its estimation are pivotal in risk management for many geophysical applications. Return period is usually estimated by inferring a probability distribution from an observed series of the random process of interest and then applying the classical equation, i.e. the inverse of the exceedance probability. Traditionally, we form a statistical sample by selecting, from the "complete" time series (e.g. at the daily scale), those values that can reasonably be considered as realizations of independent extremes, e.g. annual maxima or peaks over a certain high threshold. Such a selection procedure entails that a large number of observations are discarded; this wastage of information could have important consequences in practical prob-

*Corresponding author
*Email address:* elena.volpi@uniroma3.it (Elena Volpi)

lems, where the reduction of the already small size of common hydrological records significantly affects the reliability of the estimates. Under such circumstances, it is crucial to exploit all the available information. To this end, we investigate the advantages of estimating the return period without any data decimation, by using the full data-set. The proposed procedure, denoted as *Complete Time-series Analysis* (CTA), exploits the property that the average interarrival time (i.e. return period) of potentially damaging events is not affected by the dependence structure of the underlying process, even for cyclo-stationary (e.g. seasonal) processes. For the sake of illustration, the CTA is compared to that based on annual maxima selection, through a simple non-parametric approach, discussing advantages and limitations of the method. Results suggest that the proposed CTA approach provides a more conservative return period estimation in an holistic implementation framework within a broader range of return period values than that pertaining to other methods, which means not only the largest extremes that are the focus of extreme value theory.

*Keywords:* Return period; Interarrival time; Complete Time-series Analysis; Persistence; Seasonality; Annual maxima

## 1. Introduction

Hydraulic risk analysis relies on finding the probability of failure of a given hydraulic structure or, more generally, system due to the occurrence of intense hydrological events, where the probability of failure is usually expressed in terms of return period. Different failure mechanisms could be considered, where each of them results from the combination of multiple characteristics of the hydrological loads (Schumann, 2017). Hence, under general conditions, the return period of structure failure should be quantified taking into account the joint probability of failure mechanisms, i.e. the joint probability of the random variables describing the hydrological load and the complex interactions between the structure and the hydrological loads acting on it (see, e.g. Volpi and Fiori, 2014).

In the simplest case, we have a single failure mechanism that is ruled by a single random variable describing the hydrological load, for example a bridge destroyed by a flood. Under such circumstances, the return period of structure failure corresponds to that of the hydrological load. Once the key variable representing the hydrological load is identified, the problem is solved by inferring a probability distribution from a series of realizations of this random variable, in order to determine the magnitude of the event, corresponding to a given return period or probability of failure.

Given this premise, it is clear that the concept of return period and how it is estimated from observations is central to risk management problems in hydrological/hydraulic applications; yet this is true also in many other geophysical and engineering fields. Even if return period is a widely applied and well established probabilistic tool for hydrological applications, since the pioneering work of Alexander (1959) there have been few studies attempting to analyze the differences between estimated return periods of hydrological extremes using different methods of estimation. Nonetheless, some researchers have recently investigated the concept of return period when the basic assumptions of stationarity and independence are omitted; see, among others, Rootzén and Katz (2013); Obeysekera and Salas (2016); Read and

Vogel (2016); Fernández and Salas (1999); Douglas et al. (2002); Bunde et al. (2003); Eichner et al. (2011); Volpi et al. (2015) and references therein. As detailed in the following, the purpose of our work is to investigate a new statistical approach to infer return period from a complete record of observed data; therefore, we must assume a dependence structure in time and a stationary framework, because the non-stationary hypothesis implies a priori attributions, supported by deductive reasoning, that go beyond the scope of this paper (Koutsoyiannis and Montanari, 2015; Serinaldi et al., 2018).

Under the assumption of stationarity, Bunde et al. (2003) and Volpi et al. (2015) have shown that the independence condition is not necessary in order to apply the classical equation of return period, i.e., the inverse of the exceedance probability. Following Volpi et al. (2015), this paper highlights how temporal dependence does not alter the average interarrival time formulation, even in stochastic processes characterized by *cyclo-stationarity*, a characteristic that hydrological and other geophysical processes exhibit at sub-annual scales. Furthermore, we investigate here the potential application of this important property of return period, derived from the full available record, for frequency analysis; specifically, we show how the return period can be directly estimated from raw data records of a time-dependent process, regardless of its dependence structure, under stationary or cyclo-stationary conditions.

This alternative approach for return period estimation, which is proposed here for the first time and denoted as *Complete Time-series Analysis* (CTA), is compared to the traditional approach based on frequency analysis of Annual Maxima (AM), which constitutes the basis of traditional extreme value analysis. Indeed, we usually analyze AM to catch the tail of the distribution of the parent process, where the latter is the process of interest. Hence, the rationale behind CTA is to exploit all the information provided by observational data (Marani and Ignaccolo, 2015; Zorzetto et al., 2016), with the objective of better estimating the return period in a wider range of values, not only at the largest extremes that are the focus of extreme value theory. Note, indeed, that small to moderate return period values are still of interest in several practical problems, such as pluvial flooding. Furthermore, it is important to stress that CTA provides different return period estimates with respect to annual maxima by considering *all* the occurrences of the *dangerous* values (e.g. exceedance of the random variable above any threshold value of interest) within the observed record, as it will be discussed later on.

Hence, we aim at exploring the potential conveniences of CTA compared to traditional approaches, and not to elaborate on the return period concept (see Volpi et al., 2015). For this reason, in this article we base and limit our investigation to the non-parametric approach. Two illustrative examples are presented, both relying on synthetic processes: the first one makes use of a very simple process whose correlation structure is known a priori; the second example resembles the main characteristics of a real-word process. Future research will focus on the parametric implementation of CTA for its practical use in real world cases; indeed, the problem of fitting a model to the complete record of observations and the evaluation of the related uncertainty deserves further attention.

The remainder of this paper is organized as follows. Section 2 briefly recalls the definitions of return period available in the literature and illustrates the properties of the interarrival time for stationary and cyclo-stationary processes. In Section 3 the new approach for return period estimation, relying on frequency analysis of complete times-series (CTA), is introduced as an alternative to traditional methods based on annual maxima and peaks over threshold (Section 4). In sections 5 and 6, complete time-series analysis is compared to the standard approach using annual maxima for a simple stationary process and for a cyclo-stationary process, that mimics the characteristics of a real world phenomenon, respectively. Section 7 is concerned with the potential issues related to the application of the proposed approach in real word problems, while the Conclusion Section summarizes the main findings of this paper.

## 2. Return period: Definitions and properties

Let $Z(\tau)$ be a stochastic process that represents a natural process evolving in continuous time $\tau$. As observations of $Z$ are only made in discrete time, we consider here the corresponding discrete-time process $Z_j$ that is obtained by sampling $Z(\tau)$ at constant time intervals $\Delta\tau$, i.e. $Z_j = Z(j\Delta\tau)$ where $j$ $(= 1, 2, ...)$ denotes discrete time. We make the only assumption that $Z_j$ is a stationary process, fully described in terms of its marginal probability function $P_Z(z) = \Pr\{Z \le z\}$ and, up to the second order in terms of joint distribution, by its autocorrelation function $\rho_\theta = \gamma_\theta/\gamma_0$ (with $\theta = 0, \pm 1, \pm 2, ...$), where $\gamma_\theta = \mathrm{cov}[Z_j, Z_{j+\theta}]$ and $\rho_\theta \in [-1, 1]$; further, we denote the mean of the process as $\mu = \mathrm{E}[Z_j]$ and its standard deviation $\sigma = \sqrt{\gamma_0}$.

For design and risk assessment purposes, we are interested in the occurrence of dangerous events that might result in a system or structure failure.

We define here a dangerous event as the exceedance at the $\Delta\tau$ scale of a threshold level, $A = \{Z > z\}$, for instance the discharge exceeding a given high threshold level, potentially causing the flooding of an urbanized area. The probability of $A$ is given by $\Pr A = \Pr\{Z > z\} = 1 - P_Z(z) = 1 - \Pr B$, where $B$ denotes the complementary event of $A$. In hydrological applications, as well as in many other engineering fields, the rareness of the dangerous events is usually measured in terms of return period $T(z)$, thus assuming that the event $A$ will occur on average once every $T$ years, which is the time unit commonly used for return periods in hydrology. Mathematically, it is

$$\frac{T}{\Delta\tau} = \mathrm{E}[X] = \sum_{x=1}^{\infty} x f_X(x) \tag{1}$$

where $X$ is the number of discrete time steps to an occurrence of the event $A$ and $f_X(x) = \Pr\{X = x\}$ is its probability mass function (pmf). Note that in eq. (1), which only refers to discrete-time processes, $T$ is measured in units of time, i.e. $\Delta\tau$; if $\Delta\tau = 1$ year the return period is measured in years.

As highlighted by previous literature studies (see, e.g. Fernández and Salas, 1999; Douglas et al., 2002), the return period can be defined as the average of (i) the *waiting time*, that is the time interval ranging from the present to the next threshold exceedance, or (ii) the *interarrival time*, that is the time elapsing between any two successive realizations of the dangerous event. As explained later on, we adopt here the second definition, which implicitly assumes that a dangerous event has just occurred. We remark that such a definition is customary in hydrological applications (see, e.g. Chow et al., 1988; Kottegoda and Rosso, 1997; Salvadori et al., 2007). For convenience, herein we express discrete time as $t = j - j_0$, where $j_0$ is the current instant of time (when the dangerous event has just occurred); therefore, the discrete-time process is indicated as $Z_t$, and $t = 0$ denotes the present. As a consequence, the pmf of the interarrival time can be written as (Fernández and Salas, 1999)

$$\begin{aligned} f_X(x) &= \Pr(B_1, B_2, ...B_{x-1}, A_x | A_0) \\ &= \frac{\Pr(A_0, B_1, B_2, ...B_{x-1}, A_x)}{\Pr A_0} \end{aligned} \tag{2}$$

Then, the definition of return period based on the concept of interarrival time relies on a conditional probability. The average interarrival time is obtained by substituting Equation (2) into (1).

If $Z_t$ is a purely random process, then the return period $T$ is given by (e.g. Stedinger et al., 1993)

$$\frac{T(z)}{\Delta\tau} = \frac{1}{1 - P_Z(z)} \tag{3}$$

regardless of the definition used for $X$ in Eq.(1), i.e. waiting time or interarrival time. The above relationship holds true even if the stationary and independent process is not sampled at constant time intervals; in this case $\Delta\tau$ is the average time interval between consecutive samples (Koutsoyiannis, 2008).

Although the independence condition is typically assumed as a necessary condition for Equation (3), it has been recently demonstrated by Volpi et al. (2015) - independently from the conceptual arguments presented by Bunde et al. (2003) - that the return period $T(z)$, defined as the average interarrival time, is expressed by Eq. (3) even for processes correlated in time, with any type of dependence structure of $Z$, an important and not very well-known fact that is exploited in the following development.

Conversely, when based on the concept of waiting time, the formulation of the return period strictly depends on the correlation structure of the process; specifically, for any dangerous events $A$ (or threshold levels $z$), the average waiting time is an increasing function of the correlation of the process, thus resulting in values larger than the average interarrival time (which ignores correlation). Hence, for processes which are positively correlated in time (such as most hydrological processes) Equation (3) returns a *lower bound* in terms of $T(z)$.

Although the average interarrival time $T(z)$ remains the same for correlated and independent processes, the probability that the threshold $z$ is exceeded in a given period can be very different in the two cases. In fact, if a dangerous event occurs at present time, then the conditional probability of occurrence of another dangerous event at successive instants of time will be greater than the independent case; this yields that the probability mass function of the interarrival time (that corresponds to the probability of failure) will have a larger mass for small temporal values and a lower mass elsewhere, hence a larger variance with respect to the independent case. Since the average waiting time is an increasing function of the variance of the interarrival time, as shown in Volpi et al. (2015), the latter characteristic of the interarrival time distribution explains why the average waiting time is larger than the average interarrival time. As a further consequence, the definition of return period based on the interarrival time might result in higher values of the probability of failure with respect to the independent case; the reader is referred to Volpi et al. (2015) for further details on the theoretical properties of both definitions of $T$.

### 2.1. Cyclo-stationary processes

The property of interarrival time mentioned above has been derived based on the assumption that the underlying process $Z$ is stationary (see the Appendix B in Volpi et al. (2015) for details). However, many natural processes exhibit statistical properties that are invariant to a shift of the time origin by integral multiples of a certain period $\Pi$, due to e.g. the seasonal variability of environmental phenomena at sub-annual scales (Koutsoyiannis, 2016). In stochastic hydrology, such processes are usually modelled by cyclo-stationary processes with period $\Pi$.

Let us consider a cyclo-stationary process that is characterized by a joint distribution function that varies within the time period $\Pi$ (typically equal to one year), such that $\Pr B_t = \Pr B_{\Pi+t}$, $\Pr(B_t, B_{t+1}) = \Pr(B_{\Pi+t}, B_{\Pi+t+1})$ and so on. For such a process, the pmf given in Equation (2) can be regarded as the pmf of the interarrival time conditional to the occurrence of the dangerous event at time $t = 0$, i.e. $f_X(x|t = 0)$. For any value of $t$, this conditional pmf can be written as

$$f_X(x|t) = \frac{\Pr(A_t, B_{t+1}, B_{t+2}, ...B_{t+x-1}, A_{t+x})}{\Pr A_t} \tag{4}$$

To account for the possible occurrence of $A$ at every instant of time $t$ within the period $\Pi$, we marginalize the above conditional probability by summing the pmf in Eq.(4) with respect to all the possible values of time $t' \in [t, t+\Pi-1]$ according to their probability of occurrence. The latter quantity is nothing else than the conditional probability $\Pr(A_{t'}|t' \in [t, t+\Pi-1])$. Hence, it is

$$f_X(x|t) = \sum_{t'=t}^{t+\Pi-1} \Pr(A_{t'}|t' \in [t, t+\Pi-1]) \frac{\Pr(A_{t'}, B_{t'+1}, B_{t'+2}, ...B_{t'+x-1}, A_{t'+x})}{\Pr A_{t'}}$$
$$= \sum_{t'=t}^{t+\Pi-1} \frac{\Pr A_{t'}}{\sum_{r=t}^{t+\Pi-1} \Pr A_r} \frac{\Pr(A_{t'}, B_{t'+1}, B_{t'+2}, ...B_{t'+x-1}, A_{t'+x})}{\Pr A_{t'}} \tag{5}$$

Finally, the average interarrival time of the cyclo-stationary process is obtained by substituting in Equation (1) the pmf given in Eq. (5), thus obtaining

$$\frac{T(z)}{\Delta\tau} = \frac{1}{1 - \overline{P_Z(z)}} \tag{6}$$

where $\overline{P_Z(z)} = \frac{1}{\sum_{r=t}^{t+\Pi-1} \Pr A_r} = \frac{1}{1-\sum_{r=t}^{t+\Pi-1} \Pr B_r}$ is the marginal probability of non-exceeding the threshold value $z$ within any period $[t, t+\Pi-1]$; since we are dealing with a cyclo-stationary processes, $\overline{P_Z(z)}$ remains the same for any $t$. The derivation of Equation (6) is given in Appendix A. While for cyclo-stationary processes the exceeding probability $\Pr A_{t'} = 1 - \Pr B_{t'}$ varies with time $t' \in [t, t+\Pi-1]$, the return period of the dangerous event $A$ is a constant value, independent of time $t$, and it is expressed again by the classical equation of return period.

## 3. Novel return period estimation: Complete Time-series Analysis (CTA)

The property that the average interarrival time (i.e. return period) is not affected by the dependence structure of the underlying process, even for cyclo-stationary processes, sheds a new light on the problem of return period estimation in practical problems. Under the hypothesis of stationarity, or cyclo-stationarity in the case of processes exhibiting seasonality and sampled at the sub-annual scale, the return period can be estimated by using Equation (3) starting from any kind of observational data, independent or correlated in time, thereby potentially exploiting all the available information on the underlying process. Hence, the only necessary assumption is that of stationarity; but stationarity is also related to ergodicity, which in turn is a prerequisite to make inference from data. As previously mentioned, the stationarity issue goes beyond the scope of this work; the interested readers are referred to the work of Koutsoyiannis (2014), Montanari and Koutsoyiannis (2014), Serinaldi and Kilsby (2015), Koutsoyiannis (2016), Serinaldi et al. (2018) and Luke et al. (2017) for a comprehensive discussion.

Let $\mathbf{z} = z_1, ..., z_n$ be an observed realization of the stochastic process $Z_t$, where $n$ is the length of the data series. Following its formal definition based on the interarrival time, the empirical return period could be estimated by first deriving from $\mathbf{z}$ the sample of the interarrival time for each value of the threshold $z$, i.e. $\mathbf{x}(z) = x_1, x_2, ..., x_{\eta(z)}$, and then averaging in time, as in Equation (1). This approach can be applied only in the case of very long time-series (very large $n$), for which the size of the interarrival series $\eta(z)$ is large enough to return reliable estimates for high threshold values $z$.

In hydrological applications the length of the observed series is usually less than one hundred years (e.g. at the daily or at the hourly scale); hence it is important to exploit all the available information for sample estimates. The empirical return period can be obtained by directly applying Equation (3) to the entire time-series $\mathbf{z}$, where the cumulative probability function $P_Z$ is substituted by a probability model inferred from the observed data. For simplicity, we adopt here a non-parametric approach, by substituting $P_Z$ with its empirical counterpart. The empirical distribution function (edf), denoted in the following as $\hat{F}_n$, provides an estimate of the distribution function of the underlying stochastic process (e.g. Kolmogorov, 1933). Let $\{Z_i\}_{i=1}^n$ be a sequence of dependent and equally distributed random variables, then $F_n$ is defined as the fraction of random variables that are less than or equal to the specified value $z$, i.e.

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i \leq z\}} \tag{7}$$

where $I_{Z_i \leq z}$ is the indicator of the event $\{Z_i \leq z\}$; the estimate, $\hat{F}_n(z)$, is obtained by considering in Eq. (7) the outcome $\mathbf{z}$ instead of the random variables.

The edf in Eq. (7) is an unbiased estimator of the marginal probability function, i.e. $\mathrm{E}[F_n(z)] = P_Z(z)$; this means that the dependence structure of the underlying process does not affect the expectation of the edf, while it affects its covariance, as reported by Azriel and Schwartzman (2015). Moreover, these Authors remark that the estimator in Eq. (7) is consistent for all Gaussian stationary ergodic processes, that are characterized by an autocorrelation function decreasing to zero as the lag-time goes to infinity (i.e. the necessary condition for ergodicity). The latter case includes both short and long-range dependent processes, such as the Hurst-Kolmogorov process (Koutsoyiannis, 2016). The consistency property of the edf continues to hold for non-Gaussian distributions under various forms of dependence (see, e.g., Dedecker and Merlevéde, 2007; Wu, 2006). Note that in the case of cyclostationary processes, where the marginal probability distribution changes with time within the period $\Pi$, $\hat{F}_n$ directly provides an estimate of $\overline{P_Z(z)}$, which takes into account the variability during the period of the marginal probability function (i.e. the alternation of events characterized by a different probability of exceedance).

The computational approach based on the complete edf (complete time-series analysis, CTA), returns on average the same results to those based on the observed interarrival time-series, provided that $n$ (hence $\eta(z)$) is very large. An illustration example showing the latter property will be presented in the following sections.

The estimation approach proposed in this Section makes use of the whole available information on the process $Z$, i.e. the complete time-series $\mathbf{z}$, with the aim of returning a reliable and robust estimate of the return period of the dangerous event $\{Z > z\}$ for any threshold value $z$, i.e. in a broader range of values with respect to the traditional approaches used in extreme value analysis. It is important to stress that CTA requires the availability of an uninterrupted record of observations (i.e. $\mathbf{z}$) where the process is sampled at constant time intervals ($\Delta\tau$); indeed, this is a necessary condition for the property of the average interarrival time to hold true, as explained in Volpi et al. (2015).

In the following sections, we compare CTA with the traditional sampling methods used in extreme value analysis, thus highlighting advantages and limitations of both the strategies.

## 4. Traditional approaches for return period estimation

Since independence of $Z_t$ is usually invoked for the derivation of Equation (3) (e.g. Benjamin and Cornell (1970), p. 233, Kottegoda and Rosso (1997),

p. 190 and Chow et al. (1988), p. 383), it is common practice in hydrological applications to implement some techniques for data selection aimed to allow the assumption of the statistical independence of the observations. These techniques constitute the basis for the extreme value analysis, whose objective is to quantify the stochastic behavior of a process at unusually large or small levels that potentially lead to the failure of a system (Salvadori et al., 2007; Coles, 2001). Classical methods for extreme (and independent) value selection are the block maxima approach, where the block generally coincides with the year (Annual Maxima, AM), and the more complex Peak-Over-Threshold approach (POT).

The wide popularity of AM relies on its simplicity, but also on the limited access in the past to regularly-sampled, long observed series of the random variable of interest. Although it is necessary to have available the complete series to establish if an extreme event is an annual maximum, it was common practice to take note essentially of the values reached during extreme events, especially in old times when the observation of the hydrological variables was not systematic (apart from a few noteworthy cases such as that described in Calenda et al. (2005)).

When a complete time-series of observations is available, POT is adopted in practical applications (especially when the length of the observed period is short) to select the largest possible amount of data, yet respecting the independence assumption. POT originated in hydrology with the rationale that if additional information about the extreme upper tail were used besides the annual maxima, then more accurate and reliable estimates of the parameters and quantiles of extreme value distributions would be obtained (see, e.g.

Katz et al., 2002). This is the same rationale behind CTA, as described in the previous section; however, it is important to stress that while CTA does not require any data selection, POT asks for additional computational efforts to select only the peaks over the threshold (i.e. the maximum of a cluster of values all exceeding the threshold) and necessitates the introduction of further information or parameters in order to select among those peaks only independent ones (see, e.g. Coles, 2001).

For simplicity, in this work we compare CTA to AM. To avoid confusion, we use $Y$ to indicate the annual maximum of the random variable $Z$. The annual maximum time-series is derived from $\mathbf{z}$ as $\mathbf{y} = \{y_1, ..., y_{n/n_Y}\}$, where $n$ is the number of observations, $y_i = \max\{z_{n_Y(i-1)+1}, ..., z_{n_Y i}\}$ and $n_Y$ is the number of time-intervals $\Delta\tau$ per year (e.g. $n_Y = 365$ in the case $Z$ is observed at the daily scale). The empirical return period of the values in $\mathbf{y}$ can be evaluated by using the same procedures described above for the time-series $\mathbf{z}$ and based on the edf (since $(n/n_Y) \leq n$).

### *4.1. Purely random and stationary processes*

The probability distribution function of $Y$ is by definition different from that of $Z$. Since we aim at exploring the difference between the two under general conditions (cyclo-stationarity and persistence), for the sake of clarity we start from the well known stationary and independent case (then $\rho_\theta = 0$ for $\theta \neq 0$) by introducing a general framework that is instrumental to the discussion reported in the following sections. Given a threshold value $z$, the probability of annual maxima $P_Y(z) = \Pr\{Y \leq z\}$ can be easily derived from that of the parent process as (e.g. Coles, 2001)

$$P_Y(z) = P_Z(z)^{n_Y} \tag{8}$$

By using Equation (3) the corresponding return period is derived

$$T_Y(z) = \frac{\Delta\tau\ n_Y}{1 - P_Z(z)^{n_Y}} \tag{9}$$

where $\Delta\tau\ n_Y$ is one year. Note that here $n_Y$ is not a random variable, since $\mathbf{z}$ is an observed series of the stochastic process $Z$ sampled at a constant time intervals $\Delta\tau$; hence, the number of observations in each year is constant, being uniquely determined by the time interval $\Delta\tau$. Conversely, in traditional extreme value theory the exponent in Eq. (8) is not a constant, being the number of peaks of clusters of values, but rather can be regarded as a realization of a Poisson distributed random variable; this yields a different form for the probability distribution of annual maxima, which gives numerical values not significantly different from those provided by Eq. (8) for large $P_Z(z)$ (Koutsoyiannis, 2004).

If $n_Y = 1$ CTA obviously gives the same results of AM. If $n_Y > 1$ (which means that $\Delta\tau < 1$ year), Equation (9) results in larger values with respect to $1/(1 - P_Z(z))$, as shown in Figure 1. Note that the figure depicts the theoretical return period of annual maxima as function of that of the continuous process, i.e. when assuming $n_Y \to \infty$, for any kind of process $Z$; in other words, it is the case of infinite sample length ($n \to \infty$). For convenience, the return period of annual maxima is denoted by $T_Y$ while that of the parent process by $T_Z$; both are measured in years from now on. The figure shows to what extent AM results in larger values of the return period with respect to that of the underlying process, or in other words, the probability of $\{Y > z\}$ is smaller than that of the dangerous events $\{Z > z\}$. This is due to a *wastage* of information. During any year, additional events may have occurred that are excluded by the analysis because such data are not the annual maximum in the year they arose, as in the example depicted in Figure 2a. For the sake of illustration, Figure 2 shows one year of a log-normal AR(1) daily time-series with mean $\mu = 1$, variance $\sigma^2 = 1$ and lag-1 correlation coefficient $\rho_1 = 0$ (panel a) and $\rho_1 = 0.85$ (panel b); note that seasonal variability is not considered in this independent and stationary example. The figure depicts with red dots the information discarded by the annual maxima approach (red circles), in the estimation of the return period of the event $Z > 4$. Furthermore, these events might be possibly larger than the maximum in other years. It also follows that the minimum value of the return period of annual maxima is equal to 1 year, which means that based on annual maxima analysis we cannot measure the rareness of events that occur more frequently than once every year.

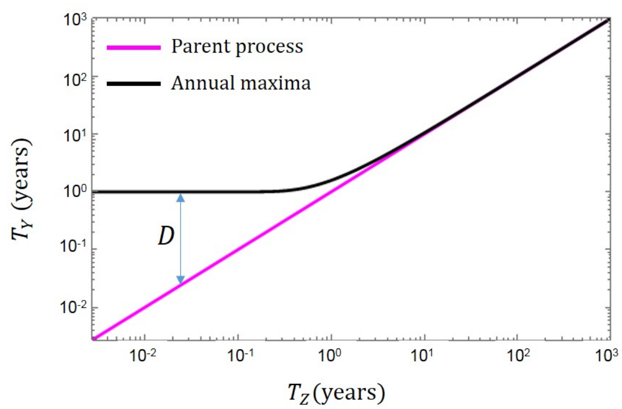


Figure 1: Return period of the annual maxima ($T_Y$) as function of that of the continuous parent process ($T_Z$, $n_Y \to \infty$). The difference between the two ($D$) significantly reduces ($D \leq 0.05$) only when when $T_Z$ becomes larger than about 10 years.

As the threshold $z$ increases (we look to more and more intense events), the return period of annual maxima, $T_Y$, tends to that of the parent process,

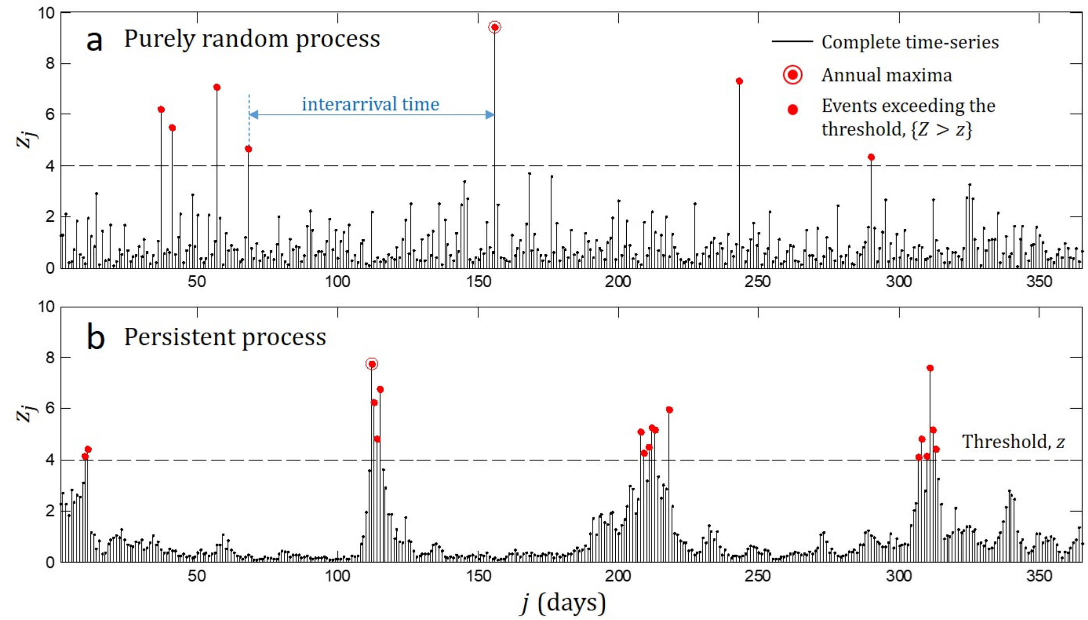


Figure 2: One year of a lognormal AR(1) daily time-series with mean $\mu = 1$, variance $\sigma^2 = 1$ and lag-1 correlation coefficient a) $\rho_1 = 0$ and b) $\rho_1 = 0.85$ (no seasonal variability). The events exceeding the threshold (red dots) that are not annual maxima (red circle) are discarded in AM, resulting in an overestimation of the average interarrival time of the parent process, i.e. $T_Y(z) > T_z(z)$.

$T_Z$; indeed, very large events are expected to be always selected as annual maxima, hence the wastage of information is reduced toward zero as the threshold increases. The difference $T_Y - T_Z$ (denoted in Figure 1 as $D$) reduces to less than 5% of $T_Z$ only when $T_Z$ becomes larger than about 10 years. This means that relatively frequent events, that might be of interest when the expected damage is modest, are generally underestimated when only the annual maxima are available. The asymptotic convergence of the annual maxima distribution to that of the parent process for high threshold values, about $T_Z \geq 10$ years according to the above figure, is at the basis of extreme value theory.

## 5. Persistent stationary processes

The difference between annual maxima and the parent process for those events that are characterized by small to medium values of the return period is expected to worsen in the case of time-dependent, positively correlated processes, where the dangerous events tend to occur in clusters; see the example shown in Figure 2b, where the number of events neglected by annual maxima approach (the red dots) increases with respect to the independent case depicted figure 2a. The latter condition is that usually matched in hydrological applications; for instance, the rainfall amount observed at given time-scale exhibits a complex, persistent behavior in time, affected by seasonality, which depends on the time-scale itself. Hence, we compare here the empirical return period of the complete time-series to that of annual maxima for time persistent processes by making use of a simple synthetic example.

In this section we consider an autoregressive process of order one, AR(1), supposed to represent a stationary and persistent natural process observed at the daily scale (i.e. $\Delta\tau = 1$ day). The process analyzed here is characterized by a marginal lognormal probability distribution function with mean $\mu$ and variance $\sigma^2$, while its time-dependence structure is ruled by the lag-1 correlation coefficient, $\rho_1$. We assume here $\mu = 1$ and two different values for the variance, $\sigma^2 = 1$ or $\sigma^2 = 5$; further, we let the correlation coefficient vary between 0 (independent process) and 0.99 (persistent process), noting that we are ignoring seasonality here.

### 5.1. Theoretical difference between CTA and AM for persistent processes

We first compare the empirical return period of the annual maxima to that of the complete series by analyzing a very long series, specifically $n = 365 \times 10^5$ days (i.e. $10^5$ years). This analysis is intended to investigate the theoretical difference between CTA and AM for persistent processes, when the accuracy of the return period estimate is essentially not affected by the length of the observed period (as in the independent case discussed in Section 4.1). Results are represented in Figure 3 for several values of $\rho_1$ ranging between 0.5 and 0.99 and for $\sigma^2 = 1$ (Figure 3a) and $\sigma^2 = 5$ (Figure 3b); the independent case ($\rho_1 = 0$) is reported as a reference. Values of $\rho_1 \in [0, 0.5]$ are not considered since the difference with the independent case is negligible; finally, we look at results for return period values included between 1 day, that is the the minimum value that can be explored based on the temporal resolution of the available series, and 1000 years, such that estimates are not influenced by the finite length of the simulated series.

We recall here that the theoretical return period (according to Equation (3)) of the parent process is fully determined by the lognormal probability distribution, which is represented in Figure 3 by the magenta curve. The return period estimated from the complete series returns for any $\rho_1$ the theoretical distribution (black dashed curves that overlap for all $\rho_1$ the magenta curve); thus, it is not affected at all by the correlation structure of the process (as demonstrated by Volpi et al. (2015)).

Since in this numerical experiment $n$ is very large, also the empirical return period computed as the average of the interarrival time between successive events $\{Z > z\}$, i.e. by strictly following Equation (1), returns the

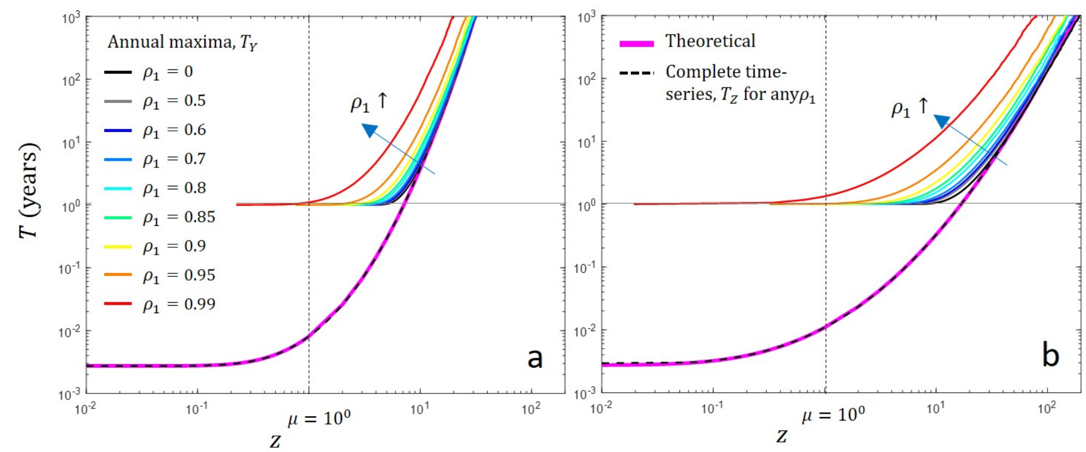


Figure 3: Lognormal AR(1) daily process with $\mu = 1$, variance $\sigma^2 = 1$ (a) or $\sigma^2 = 5$ (b), and lag-1 correlation coefficient $\rho_1$: empirical return periods of the complete series ($T_Z$, black dashed curves indistinguishable from the theoretical, thick magenta curves) and of the annual maxima ($T_Y$, colored continuous curves) for several values of $\rho_1$ ranging between 0.5 and 0.99 compared to the theoretical one (magenta curve). The independent case ($\rho_1 = 0$) is depicted as a reference.

theoretical value for any $\rho_1$ and for threshold values $z$ up to that represented in the figure. The latter result, which numerically demonstrates the theoretical finding by Volpi et al. (2015), is illustrated in Figure 4 for the specific process characterized by the parameter combination $\sigma^2 = 5$ and $\rho_1 = 0.85$.

In Figure 3, the return period estimated by selecting the annual maxima, $T_Y$ (colored curves) assumes values larger than the theoretical ones pertaining to the parent process or the independent case ($\rho_1 = 0$, black curve); this implies that the corresponding $z$-values are smaller. We also notice that for $\rho_1 \geq 0.9$ the annual maxima span over a wide range, covering values that are even smaller than the mean of the process ($\mu = 1$, vertical dashed line

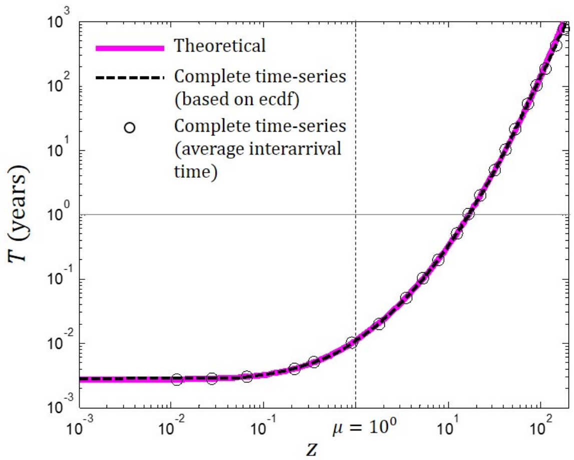


Figure 4: Lognormal AR(1) daily process with $\mu = 1$, variance $\sigma^2 = 5$ and lag-1 correlation coefficient $\rho_1 = 0.85$: empirical return period of the complete series $T_Z$, derived based on the ecdf (black dashed curves) and as the average of the empirical interarrival times (black circles) compared to the theoretical one (magenta curve).

in the figure). On average, the number of daily data exceeding the threshold $\{Z > \min(\mathbf{y})\}$ that are discarded by AM per year ranges in between 10 and 350 values when $\rho_1$ increases from 0 to 0.99 respectively, for both $\sigma^2 = 1$ and $\sigma^2 = 5$.

Further, the return period estimate converges only for large values of $z$ to the theoretical value, with a rate of convergence that depends on the persistence of the process. Hence, the larger is $\rho_1$ the slower is the rate of convergence of $T_Y$ to $T_Z$. An important role is also played by the variance of the process; in the case $\sigma^2 = 5$ (Figure 3b) the convergence of the complete distribution to the theoretical one is even slower than in the case depicted in Figure 3a ($\sigma^2 = 1$). To give a quantitative measure of the deviation of the annual maxima estimate from the theoretical return period of the parent process, the difference $D = T_Y - T_Z$ for $T_Z = 10$ years moves from 0.5 years for $\rho_1 = 0$ (i.e. $\sim 5\%$, as in the theoretical independent case depicted in Figure 1), to 150 (200) years, for $\rho_1 = 0.99$ and $\sigma^2 = 1$ (5).

Note that return period estimate based on CTA is compared here only to that pertaining to annual maxima, but a similar comparison could be made by considering the POT approach. It is expected that return period estimates based on POT result in intermediate values between AM and CTA, as a function of the threshold used to select peaks, but closer to AM estimates. Note indeed, that CTA considers all the values exceeding the threshold $z$ (see, e.g., figure 2b), while POT uses only the independent maxima of clusters of values exceeding $z$. The difference between the two approaches might be relevant for practical purposes, as discussed later in Section 7.

### *5.2. Effects of finite time-series length*

The situation depicted in Figure 3 is not met in practical applications, when the limited length of the observed series significantly affects the return period estimation in terms of both accuracy and uncertainty. Generally speaking, accuracy is expected to improve while uncertainty decreases when increasing the length of the dataset of observations of a given process; note that here the number of observations is not a direct measure of the amount of information provided by data because of the correlation among the observed values in complete time-series. Hence, we aim at comparing the overall robustness of return period estimates for small sample lengths, obtained by CTA instead of the selected annual maxima (which are commonly assumed to be independent). To investigate the latter issue, we repeated the above analysis for different values of $n$ within the range from 10 to 200 years for a large number a synthetic time-series ($M = 10,000$).

Results obtained when assuming $\sigma^2 = 5$ and $\rho_1 = 0.6$ are depicted in Figure 5; the figure summarizes the empirical probability estimates together with their 95% uncertainty bounds derived by using both methods for some values of $n$ ranging in between 10 and 200 years. Note that results are presented here in terms of edf to avoid infinite values of return period that might occur due to the use of Equation (1) when $z$ is larger than the maximum observed value in the dataset. If the edf in Equation (7) is normalized with respect to $n+1$ instead of $n$, we obtain the classic Weibull plotting position formula (Makkonen, 2006); indeed, the latter is usually adopted to avoid infinite values of the estimated return period for the sample maximum. This issue goes beyond the scope of this analysis, which is intended to discuss the variability of return period estimate due to finite sample lengths; it will be considered in future works together with the problem of model fitting.

Figure 5 shows how the AM estimate converges on the average only with increasing $T$ (moving from Figure 5a to c) to that of the complete time-series, which overlaps the theoretical one for any sample length $n$ . As expected, the uncertainty bounds reduce with the sample size $n$. Uncertainty also reduces as $T$ increases; however, this unexpected behavior is a consequence of the fact that probability is upper bounded to unity and due to the adoption of the edf given in Eq. (7). Notwithstanding this, it is worth noting that AM uncertainty bounds are narrower than those pertaining to CTA for any value of $T$. This means that the selection of annual maxima results in an undersampling effect, that manifests itself in terms of underestimation of the exceeding probability of the parent process (i.e. overestimation of the non-exceeding probability or of the return period as discussed in Figure 3), and of its sampling variability.

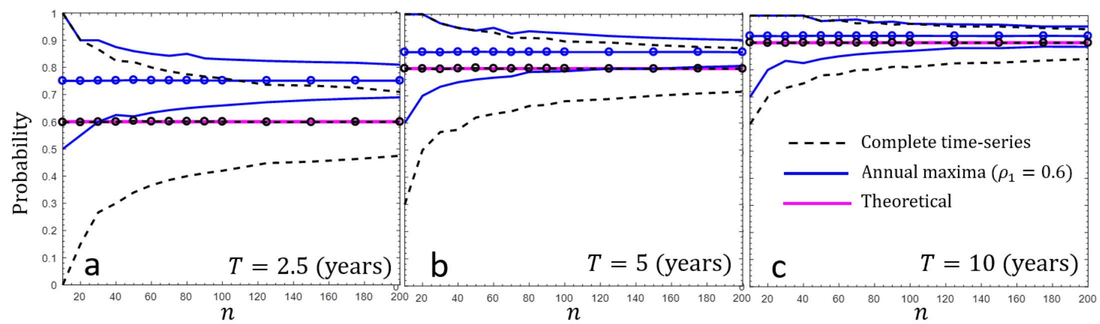


Figure 5: Lognormal AR(1) daily process with $\mu = 1$, variance $\sigma^2 = 5$ and lag-1 correlation coefficient $\rho_1 = 0.6$: CTA (black curves) and AM (blue curves) empirical probability for some values of the return period $T$ as function of the time-series length, $n$; the estimated probabilities are represented in terms of average values (dotted curve) and 95% uncertainty bounds. In each panel the theoretical probability is reported as a reference (magenta curve).

Correlation in time, which in this case is fully represented by the lag-1 correlation coefficient $\rho_1$, significantly affects the accuracy and the variability of the return period estimates obtained by both methods. Results for all values of $\rho_1$ considered in this illustrative example, are summarized in Figure 6 for $T = 5$ years. Figure 6a and b depict the behavior of the average probability estimates based on AM (panel a) and CTA (panel b). It can be noticed that the *bias* resulting from AM is strongly enhanced by high values of $\rho_1$. Conversely, the average probability estimates based on CTA are unbiased for any $\rho_1$ and $n$.

Further, the underestimation of the sampling variability which is observed in Figure 5 for $\rho_1 = 0.6$ when using AM, magnifies in the cases of processes strongly correlated in time. To illustrate the latter issue we depict in Figure 6 also the coefficient of variation, $C_V$ of the probability estimates for $T = 5$ years, computed by analyzing annual maxima (Figure 6c) and the complete time-series (Figure 6d). The figure clearly shows that while $C_V$ is comparable for small values of $\rho_1$, large differences arise when $\rho_1$ approaches to one. While $C_V$ of CTA estimate increases with the persistence of the process, that of AM reduces; the latter behaviour is a consequence of the fact that probability, that is upper bounded by one, is overestimated when analyzing annual maxima.

A similar analysis could be performed in terms of quantiles, by investigating how the order statistics of annual maxima and complete time-series are influenced by the correlation structure of the process. However, slightly different results (not shown) are obtained in terms of empirical return period quantiles with respect to those obtained in terms of edf (as in Figure 6). In fact, the probability distribution of the order statistics does not correspond on average to the theoretical probability distribution of the underlying process (David and Nagaraja, 2003); moreover, it is affected by the autocorrelation structure of the process. Conversely, the edf expressed by Equation (7) is an unbiased estimator regardless of the type and strength of the correlation structure.

## 6. Persistent and cyclo-stationary processes

In order to provide some insights into the use of CTA in applications, we analyze here a synthetic process which resembles the main statistical properties of an hydrological observed series. Specifically, we analyze a daily

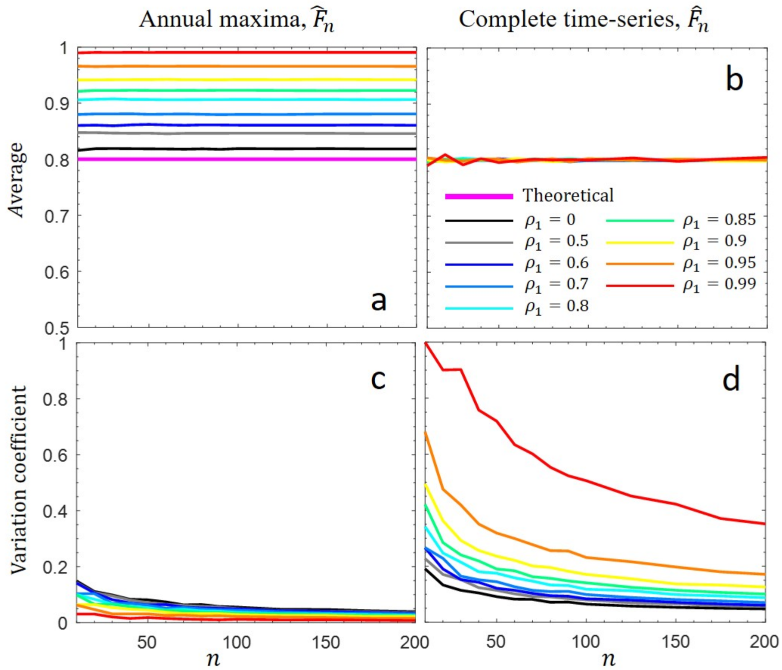


Figure 6: Lognormal AR(1) daily time-series, mean $\mu = 1$, variance $\sigma^2 = 5$, and lag-1 correlation coefficient $\rho_1 \in [0.50, 0.99]$: AM (left panels) and CTA (right panels) ecdf $\hat{F}_n$ for $T = 5$ years; average values (upper panels) and variation coefficient (lower panels) as function of the time-series length, $n$. In panels (a) and (b) the theoretical probability (magenta line) and the independent case ($\rho_1 = 0$, black curve) are reported as reference.

discharge process that is characterized by non-normality, a strong seasonal pattern and by long-range persistence.

The type of analysis envisaged here requires a very long series of data. Hence, for the sake of illustration we consider a fractional autoregressive moving average process, FARMA$(p, d, q)$, which models the Tiber River daily discharge time-series observed at Roma-Ripetta station. Observations cover a period of 54 years, from 1930 to 1983, but only the first 15 years were used to calibrate the linear parametric model; as an example, Figure 7a shows the observed series (black line) for a time window of three years, from 1933 to 1936. The seasonal pattern clearly emerges from the structure of the autocorrelation function, as depicted in Figure 7b (black line). The model was calibrated after normalizing the data (based on a log-normal transformation) and removing seasonality; Figure 7 shows a subsample of the simulated series compared to the observed one (panel a) and the corresponding autocorrelation functions (panel b), thus highlighting the capability of the model in reproducing the complex behavior of the real word process. The reader is referred to Grimaldi (2004) for a detailed description of model structure, calibration and performance. We remark again that the model employed here is for the sole sake of illustration, and other general and more parsimonious methods could have been used to generate synthetic series from the observed process with any arbitrary autocorrelation structure, as discussed by Koutsoyiannis (2016), yet this goes beyond the scope of this work.

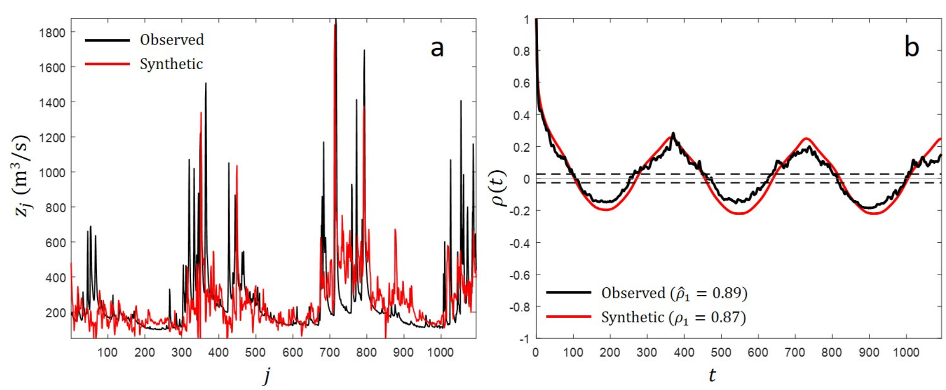


Figure 7: Synthetic series compared to that of the Tiber River (1930-1985) for a time window of three years: a) discrete-time daily discharge, and b) autocorrelation function, where dashed lines show 95% Gaussian confidence band.

We estimated the empirical return periods for both the annual maxima and the complete time-series by substituting in Eq. (3) the non-exceedance probability with the edf calculated using Equation (7), which gives the average non-exceeding probability within the year ($\Pi$). The empirical return periods are depicted in Figure 8a; since the synthetic series is very long, the figure depicts the theoretical difference between the two (unaffected by sample length). AM significantly overestimates the return period of the complete time-series, if the latter is considered as a benchmark, for return periods up to 100 years. Here the bias of AM with respect to CTA ($D = T_Y - T_Z$) for $T_Z = 10$ years is equal to $\sim$ 10 years, which means a 100% relative difference. The latter value is very close to that pertaining to the AR(1) lognormal process with similar value of $\rho_1$ and $\sigma^2 = 5$, discussed in Section 5.2, although the variation coefficient here is smaller (about 0.7) with respect to that pertaining to the AR(1) process (about 2.3).

If a finite length sample is used to estimate return periods, the difference between AM and CTA might be enhanced. The effects of finite sample length for this cyclo-stationary, long-range persistent process is illustrated in panels b)-d) of Figure 8; in panels b) and c) the empirical return periods of two samples of 54 years (equal to the length of the observed time-series) drawn from the FARMA calibrated model are compared to the corresponding *theoretical* ones (the full length samples). The empirical return period estimates for the event $\{Z > 1500 \text{ m}^3/s\}$ are depicted in panel (d); the average return period of annual maxima overestimates that pertaining to the complete time-series, also showing a smaller dispersion around its average value.

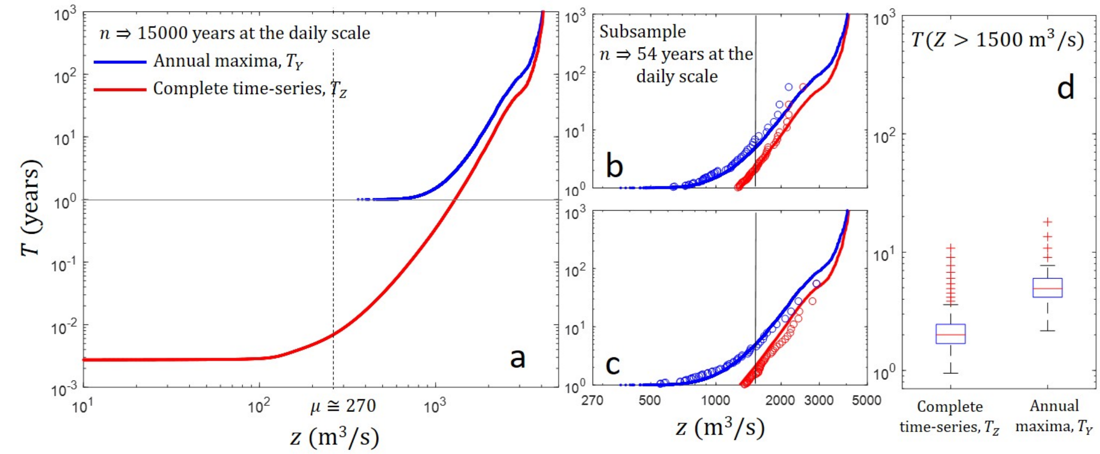


Figure 8: Synthetic daily process resembling the main statistics of the Tiber River daily discharge series (1930-1985) with mean $\mu = 270$ (m$^3/s$), standard deviation $\sigma = 181$ (m$^3/s$), lag-1 correlation coefficient $\rho_1 = 0.87$ and Hurst coefficient $H = 0.9$: empirical return period of the complete series ($T_Z$, red curve) and the annual maxima ($T_Y$, blue curve) considering the whole time-series (a) or analyzing a sub-sample of length equal to that of the observed series (b, c). Note that panels b) and c) focus on $z \geq \mu$ and $T \geq 1$ year. Panels d) shows the boxplot of the estimated return periods for the events $\{Z > 1500 \text{ m}^3/s\}$, based on 54 years sample length and both the methods.

## 7. Discussion on CTA application in real-world cases

It is important to note that the CTA approach to return period estimate considered here generally differs from the common methods used in hydrology (e.g. in flood frequency analysis), as explained in the following. CTA gives the return period of the event $A$ defined as the exceedance of a threshold value (i.e. $Z_t > z$) at the temporal resolution $\Delta\tau$ at which the continuous process $Z$ is sampled. In other words, it accounts for all the interarrival times between any successive values exceeding $z$ at the $\Delta\tau$ scale, including those elapsing between successive values of $Z_t$ that remain above the threshold $z$ (i.e. simply equal to $\Delta\tau$, see figure 2); instead, in the conventional flood analysis the above interarrival times are usually not considered. The significance of such a return period estimate depends on the particular goal at hand and on the temporal resolution $\Delta\tau$ that should be comparable (not much smaller) with respect to the temporal scale that characterizes the natural phenomenon. The temporal scale should not be confused with the characteristic scale of the correlation structure (i.e. the integral scale), which itself depends on the temporal resolution $\Delta\tau$.

For instance, in the example provided in Section 6, CTA accounts for the consecutive exceedance of any threshold value of flow discharge at the daily scale ($\Delta\tau = 1$ day), which is a small temporal scale with respect to the average duration of a flood event. In this case, if the purpose of the analysis is the assessment of the levee system, it might be not of interest to know if the levee height is exceeded the day after once exceedance is already occurred the day before. On the other hand, it might be important to account for successive exceedances at the temporal resolution of the process $\Delta\tau$, e.g. when we are evaluating the return period of daily rainfall to design a urban drainage system against pluvial flooding and the critical duration of the system (that maximizes the peak discharge) is approximately one day.

Further, as anticipated in the Introduction, for simplicity reasons we based our analyses on a non-parametric estimation approach, i.e. using the edf in Eq. (7) instead of a model fitted to the data. However, direct estimation of the statistics of a process is generally not possible merely from the data and data alone do not enable extrapolation of estimates, as often required for planning and design purposes (Koutsoyiannis, 2016). Thus, the issue of fitting appropriate models in the context of the proposed approach deserves further investigation. Although uncertainty is inherent in any statistical model, such uncertainty could be reduced by the utilization of all the available information, as well as by judicious choices of model.

The use of CTA naturally implements seasonality handling in frequency analysis. As recently discussed by Allamanno et al. (2011) (see also Rasmussen and Rosbjerg (1991) and Strupczewski et al. (2012)), disregarding seasonality in hydro-climatic extreme value analysis, based on annual maxima or POT, leads to an overestimation of return period, which is less safe. The problem is solved by taking into account the events that occurred in all the seasons by fitting different distributions to the maxima in separate seasons or months and mixing the seasonal distributions according to their probability of occurrence (see, e.g. Mascaro, 2018) or by directly including the seasonal rate of occurrence of the exceedance events in the POT approach as in Rasmussen and Rosbjerg (1991) and Allamanno et al. (2011). CTA implements the former method, by considering for frequency analysis all the observed values in each of the seasons; this could eventually require the adoption of complex probability models (e.g. mixed models).

The adoption of complex probability models could also help handle the heterogeneity due to the possible superposition of different physical processes ruling the statistical behaviour of the random variable of interest (rainfall, floods, etc.). In fact, the general understanding appears to be that low and ordinary hydrological events could be dominated by a different process (see, e.g. Merz and Blöschl, 2008), thus having little or no contribution to the larger events. This might emerge from the edf of the annual maxima, by manifesting a different statistical behavior for ordinary and extreme events; this heterogeneity is expected to emerge more strongly when analyzing the complete time-series that brings a larger number of values (also on the upper tail of the probability distribution function) with respect to annual maxima. In this regard, we believe that a priori physical knowledge about the underlying processes, if available, could be included in the analysis to support the assumption of complex mixed models for modeling (and extrapolation) purposes (see, e.g. Calenda et al., 2009). In absence of additional knowledge on the physical process, the heterogeneity assumption cannot be truly tested; however, all events generally occur under the combination of numerous factors, so that the probabilistic treatment of processes is by definition a macroscopic approach that does not care about each of the specific factors and reduces to fitting the most appropriate model to the empirical distribution of the complete data set (or to its part that is of interest for the specific problem at hand).

Finally, we remark that CTA is based on the availability of a discrete-time uninterrupted record of observations with an adequate temporal resolution, which according to our analysis affects the results. In the case of few missing data within long observational records, some (temporal or spatial) interpolation techniques could be adopted to fill the gaps; in general, large gaps could affect CTA more than AM or POT. The improvement of large datasets of environmental observations is expected to favor CTA approach in the future.

## 8. Conclusions

A new approach for return period estimation is proposed, denoted as Complete Time-series Analysis (CTA). This approach relies on the property that the average interarrival time between successive events (e.g. $\{Z > z\}$) is not affected by the correlation structure of the underlying process, regardless of the persistence of the process. This means that independence is not a necessary condition when return period is defined as the average of the interarrival time; this also implies that no data selection techniques are needed to assure independence of the data for frequency analysis. Hence, once stationarity can be assumed, the return period can be computed by using the classical equation of return period (the inverse of the exceedance probability) starting from any kind of observational data, independent or correlated in time, thereby potentially exploiting all the available information on the parent process. This property is extended herein to include cyclo-stationary processes, since hydrological and other geophysical processes typically manifest a weaker form of stationarity at sub-annual scales connected to the seasonal variability of the environmental phenomena.

We compare the proposed approach to the simple method of Annual Maxima (AM), typically adopted in extreme value analysis. Complete time-series (observed in discrete time, at constant time intervals) and annual maxima are inherently different processes, that give subtly different information on the same underlying continuous process; specifically, CTA describes the marginal behaviour of the whole parent process (sampled in discrete time at a given temporal resolution), while AM describes the statistical behavior of its extremes.

The difference between CTA and AM are discussed herein by making use of two illustrative examples. In both cases, we adopt for the sake of illustration a non-parametric approach by using the empirical probability distribution function of the annual maxima and the complete time-series as an estimate of the marginal probability distribution function of the underlying process. Some general conclusions can be drawn from the analyses, as listed in the following.

- CTA results in an accurate estimate of return period of the parent process for any intensity of the event (i.e. threshold value $z$) and, on the average, for any sample length, regardless of the correlation structure and the seasonality of the parent process, thus allowing to investigate a wider range of return period values, not only the largest extremes that are the focus of extreme value theory.

- AM leads to an overestimation of the return period (and an underestimation of its sampling variability) of the parent process for small to moderate return period values, converging to CTA estimates for large events. This behaviour, which is a consequence of data selection and is well known in the literature for independent processes, is enhanced by time-persistence of the underlying process; further, it is independent on average from the sample length.

- Return period estimates provided by CTA are generally different with respect to that pertaining to annual maxima because CTA considers all the occurrences of the dangerous events within the observed record; their significance depends on the particular goal at hand and on the temporal resolution of the process, that should be comparable with respect to the temporal scale that characterizes the natural phenomenon.

- CTA could be easily applied in the case of complex hydrological time-series (such as that discussed here that reproduces the main features of the Tiber River daily discharge as observed at Rome-Ripetta station), which are typically characterized by non-normality, seasonality, long-range persistence and, possibly, heterogeneity.

We found that the estimation of the return period using CTA could be a convenient alternative to existing methods as function of the problem at hand, for a few reasons. First, the method is easy to implement; it can be employed for any sample length, without any data selection (e.g. events selection in flood analysis). Moreover, CTA always results in more conservative return period estimates, i.e. smaller estimated values with higher uncertainty, by exploiting all the information content of the observed data, i.e. low, ordinary and extreme discharge values that make-up the complete time-series and fully describe the seasonal pattern.

The difference between CTA and AM tends toward zero as we look at events that are more and more extreme simply because very large events are expected to be always selected as annual maxima. Thus, CTA and extreme value analysis are expected to give the same results in terms of the upper tail of the distribution, thus supporting the adoption of extreme value analysis when the interest is in large return period values. Note, however, that very intense events typically pertain to the extrapolation range, where differences among the methods could emerge when a probability distribution model is fitted to the sample.

Hence, additional work is needed to fully understand advantages and limitations of CTA in engineering practice. Since the main scope of this work was to explore the potential advantages of the complete time-series approach compared to traditional ones, we have not addressed the important issue of the inference of the statistical distribution of the hydrological variable of interest. Future work will investigate the problem of fitting appropriate candidate models able to reproduce the complex, potentially heterogeneous statistical behavior of complete time-series.

## Acknowledgments

We acknowledge the Associate Editor, Geoff Pegram and two anonymous Reviewers for thoughtful comments that helped improve the original version of the manuscript. E.V. and A.F. acknowledge funding from the Italian Ministry of Education, University and Research (MIUR), in the frame of the Departments of Excellence Initiative 2018-2022, attributed to the Department of Engineering of Roma Tre University.

## Appendix A. Mean interarrival time of cyclo-stationary and persistent processes

The average interarrival time is obtained by substituting in the general expression (1) the pmf given in Eq. (5). Note that the latter pmf depends on time $t$; hence, in the following we derive the expression of the return period

$T$ conditional on $t$, i.e. $\frac{T}{\Delta\tau} = \mathrm{E}[X|t]$

$$\begin{aligned}
\frac{T}{\Delta\tau} &= \sum_{x=1}^{\infty} x \sum_{t'=t}^{t+\Pi-1} \frac{\Pr A_{t'}}{\sum_{\eta=0}^{\Pi-1} \Pr A_\eta} \frac{\Pr(A_{t'}, B_{t'+1}, B_{t'+2}, ...B_{t'+x-1}, A_{t'+x})}{\Pr A_{t'}} \\
&= \frac{1}{\sum_{r=t}^{t+\Pi-1} \Pr A_r} \sum_{t'=t}^{t+\Pi-1} \sum_{x=1}^{\infty} x \; \Pr(A_{t'}, B_{t'+1}, B_{t'+2}, ...B_{t'+x-1}, A_{t'+x}) \\
&= \frac{1}{\sum_{r=t}^{t+\Pi-1} \Pr A_r} \sum_{t'=t}^{t+\Pi-1} [1 \Pr(A_{t'}, A_{t'+1}) + 2 \Pr(A_{t'}, B_{t'+1}, A_{t'+2}) + \\
&\qquad + 3 \Pr(A_{t'}, B_{t'+1}, B_{t'+2}, A_{t'+3}) + ...]
\end{aligned} \tag{A.1}$$

By making use of the identity $\Pr(CA) = \Pr(C) - \Pr(CB)$, where $B$ always denotes the opposite event of $A$, we obtain (as in Volpi et al. (2015))

$$\begin{aligned}
\frac{T}{\Delta\tau} &= \frac{1}{\sum_{r=t}^{t+\Pi-1} \Pr A_r} \sum_{t'=t}^{t+\Pi-1} [(\Pr A_{t'} - \Pr(A_{t'}, B_{t'+1})) + 2(\Pr(A_{t'}, B_{t'+1}) - \Pr(A_{t'}, B_{t'+1}, B_{t'+2})) + ...] \\
&= \frac{1}{\sum_{r=t}^{t+\Pi-1} \Pr A_r} \sum_{t'=t}^{t+\Pi-1} [\Pr A_{t'} + \Pr(A_{t'}, B_{t'+1}) + \Pr(A'_t, B_{t'+1}, B_{t'+2}) + ...]
\end{aligned} \tag{A.2}$$

Using once more the same identity, we find

$$\begin{aligned}\frac{T}{\Delta\tau} &= \frac{1}{\sum_{r=t}^{t+\Pi-1}\Pr A_r}\sum_{t'=t}^{t+\Pi-1}[(1-\Pr B_{t'}) + (\Pr B_{t'+1} - \Pr(B_{t'}, B_{t'+1}))+ \\ &\quad + (\Pr(B_{t'+1}, B_{t'+2}) - \Pr(B_{t'}, B_{t'+1}, B_{t'+2})) + ...] \\ &= \frac{1}{\sum_{r=t}^{t+\Pi-1}\Pr A_r}[\sum_{t'=t}^{t+\Pi-1} 1 - \sum_{t'=t}^{t+\Pi-1}\Pr B_{t'} + \sum_{t=t}^{t+\Pi-1}\Pr B_{t'+1} - \sum_{t'=t}^{t+\Pi-1}\Pr(B_{t'}, B_{t'+1})+ \\ &\quad + \sum_{t'=t}^{t+\Pi-1}\Pr(B_{t'+1}, B_{t'+2}) - ...] \\ &= \frac{\Pi}{\sum_{r=t}^{t+\Pi-1}\Pr A_r} = \frac{1}{1-\frac{1}{\Pi}\sum_{r=t}^{t+\Pi-1}\Pr B_r}\end{aligned} \tag{A.3}$$

which simplifies in Eq.(6) thanks to the periodic property of the cyclo-stationary process, such that $\sum_{t'=t}^{t+\Pi-1}\Pr B_{t'} = \sum_{t'=t}^{t+\Pi-1}\Pr B_{t'+1}$, $\sum_{t'=t}^{t+\Pi-1}\Pr(B_{t'}, B_{t'+1}) = \sum_{t'=t}^{t+\Pi-1}\Pr(B_{t'+1}, B_{t'+2})$ and so on, and that marginal probability of non-exceeding the threshold value $z$ within any period $[t, t+\Pi-1]$, i.e. $\overline{P_Z(z)} = \frac{1}{1-\frac{1}{\Pi}\sum_{r=t}^{t+\Pi-1}\Pr B_r}$, is independent on $t$.

## References


A. Schumann, Flood safety versus remaining risks - Options and limitations of probabilistic concepts in flood management, Water Resources Management, 31, 2017, 3131–3145.

E. Volpi, A. Fiori, Hydraulic structures subject to bivariate hydrological loads: Return period, design, and risk assessment, Water Resources Research, 2014, 885–897.

G.N. Alexander, Return period relationships, Journal of Geophysical Research , 64, 1959, 675–982.

H. Rootzén, R. Katz, Design life level: Quantifying risk in a changing climate, Water Resources Research, 49, 2013, 5964–5972.

J. Obeysekera, J. D. Salas, Frequency of Recurrent Extremes under Nonstationarity, Journal of Hydrologic Engineering, 21, 2016.

L. K. Read, R. M. Vogel, Hazard function analysis for flood planning under nonstationarity, Water Resources Research, 2016.

D. Koutsoyiannis, A. Montanari, Negligent killing of scientific concepts: the stationarity case, Hydrological Sciences Journal, 60, 2015, 2–22.

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

F. Serinaldi, C. Kilsby, Understanding Persistence to Avoid Underestimation of Collective Flood Risk, Water, 8, 2016, 152.

H. Tyralis, P. Dimitriadis, D. Koutsoyiannis, P. O'Connel, K. Tzouka, T. Iliopoulou, On the long-range dependence properties of annual precipitation using a global network of instrumental measurements, Advances in Water Resources, 2017.

B. Fernández, J. D. Salas, Return period and risk of hydrologic events. II: Applications, Journal of Hydrologic Engineering, 4, 1999, 308–316.

E. M. Douglas, R. M. Vogel, C. N. Kroll, Impact of streamflow persistence on hydrologic design, Journal of Hydrologic Engineering, 7, 2002, 220–227.

A. Bunde, J. F. Eichner, S. Havlin, J. W. Kantelhardt, The effect of long-term correlations on the return periods of rare events, Physica A: Statistical Mechanics and its Applications 330, 2003, 1–7.

J. F. Eichner, J. W. Kantelhardt, A. Bunde, S. Havlin, The Statistics of return intervals, maxima, and centennial events under the influence of long term correlation, in: In Extremis, Springer-Verlag, Berlin Heidelberg, 2011, pp. 3–43.

E. Volpi, A. Fiori, S. Grimaldi, F. Lombardo, D. Koutsoyiannis, One hundred years of return period: Strengths and limitations, Water Resources Research, 2015, 1–16.

M. Marani, M. Ignaccolo, A metastatistical approach to rainfall extremes, Advances in Water Resources, 79, 2015, 121–126.

E. Zorzetto, G. Botter, M. Marani, On the emergence of rainfall extremes from ordinary events, Geophysical Research Letters, 2016, 8076–8082.

V. T. Chow, D. R. Maidment, L. W. Mays, Applied hydrology, McGraw-HiII, New York, 1988.

N. T. Kottegoda, R. Rosso, Probability, statistics, and reliability for civil and environmental engineers, McGraw-Hill, Milan, 1997.

G. Salvadori, C. De Michele, N. Kottegoda, R. Rosso, Extremes in Nature – An Approach Using Copulas, Springer, New York, 2007.

J. R. Stedinger, R. M. Vogel, E. Foufoula-Georgiou, Frequency analysis of extreme events, in: D. Maidment (Ed.), Handbook of Hydrology, McGraw-Hill, New York, 1993.

D. Koutsoyiannis, Probability and statistics for geophysical processes, National Technical University of Athens, Athens, 2008.

D. Koutsoyiannis, Generic and parsimonious stochastic modelling for hydrology and beyond, Hydrological Sciences Journal, 61, 2016, 225–244.

D. Koutsoyiannis, Reconciling hydrology with engineering, Hydrology Research, 45, 2014, 1174–1183.

A. Montanari, D. Koutsoyiannis, Modeling and mitigating natural hazards: Stationarity is immortal!, Water Resources Research, 50, 2014, 9748–9756.

F. Serinaldi, C. Kilsby, Stationarity is undead: Uncertainty dominates the distribution of extremes, Advances in Water Resources, 77, 2015, 17–36.

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

A. N. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, Giornale dell'Istituto Italiano degli Attuari, 4, 1933, 83–91.

D. Azriel, A. Schwartzman, The empirical distribution of a large number of correlated normal variables, Journal of the American Statistical Association, 110, 2015, 1217–1228.

J. Dedecker, F. Merlevéde, The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in $L^p$, ESAIM: Probability and Statistics, 11, 2007, 102–114.

W.B. Wu, Oscillations of empirical distribution functions under dependence, High Dimensional Probability, Institute of Mathematical Statistics, 2006, 53-61.

J. R. Benjamin, A. A. Cornell, Probability, Statistics, and Decision for Civil Engineers, McGraw-HiII, New York, 1970.

S. Coles, An introduction to statistical modeling of extreme values, Springer, London, 2001.

G. Calenda, C. Mancini, E. Volpi, Distribution of the extreme peak floods of the Tiber River from the XV century, Advances in Water Resources, 28, 2005, 615–625.

R. W. Katz, M. B. Parlange, P. Naveau, Statistics of extremes in hydrology, Advances in Water Resources, 25, 2002, 1287–1304.

D. Koutsoyiannis, Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation, Hydrological Sciences Journal, 49, 2004, 574–590.

L. Makkonen, Plotting Positions in Extreme Value Analysis, Journal of Applied Meteorology and Climatology, 45, 2006, 335–340.

H. A. David, N. H. Nagaraja, Order statistics, John Wiley & Sons, New York, 2003.

S. Grimaldi, Linear parametric models applied to daily hydrological series, Journal of Hydrologic Engineering, 9, 2004, 383–391.

P. Allamanno, F. Laio, P. Claps, Effects of disregarding seasonality on the distribution of hydrological extremes, Hydrology and Earth System Sciences, 15, 2011, 3207–3215.

P. Rasmussen, D. Rosbjerg, Prediction uncertainty in seasonal partial duration series, Water Resources Research, 27, 1991, 2875–2883.

W. Strupczewski, K. Kochanek, E. Bogdanowicz, I. Markiewicz, On seasonal approach to flood frequency modelling. Part I: Two-component distribution revisited, Hydrological Processes, 2012.

G. Mascaro, On the distributions of annual and seasonal daily rainfall extremes in central Arizona and their spatial variability, Journal of Hydrology, 559, 2018.

B. Merz, G.. Blöschl, Flood frequency hydrology: 1. Temporal, spatial, and causal expansion of information, Water Resources Research, 44(8), 2008.

G. Calenda, C. Mancini, E. Volpi, Selection of the probabilistic model of extreme floods: The case of the River Tiber in Rome, Journal of Hydrology, 371, 2009, 1–11.