General simulation algorithm for autocorrelated binary processes

Francesco Serinaldi[1,2,*] and Federico Lombardo[3,†]

[1] *School of Civil Engineering and Geosciences, Newcastle University - Newcastle Upon Tyne, NE1 7RU, UK*

[2] *Willis Research Network - 51 Lime St., London, EC3M 7DQ, UK*

[3] *Dipartimento di Ingegneria, Università degli Studi Roma Tre - Via Vito Volterra 62, 00146 Rome, Italy*

(Dated: January 24, 2017)

Abstract

The apparent ubiquity of binary random processes in physics and many other fields has attracted considerable attention from the modeling community. However, generation of binary sequences with prescribed autocorrelation is a challenging task owing to the discrete nature of the marginal distributions, which makes the application of classical spectral techniques problematic. We show that such methods can effectively be used if we focus on the parent continuous process of beta distributed transition probabilities rather than on the target binary process. This change of paradigm results in a simulation procedure effectively embedding spectrum-based iterative amplitude adjusted Fourier transform method devised for continuous processes. The proposed algorithm is fully general, requires minimal assumptions, and can easily simulate binary signals with power-law and exponentially decaying autocorrelation functions corresponding for instance to Hurst-Kolmogorov and Markov processes. An application to rainfall intermittency shows that the proposed algorithm can also simulate surrogate data preserving the empirical autocorrelation.

PACS numbers: 02.50.-r, 02.70.Hm, 05.45.Tp

* francesco.serinaldi@ncl.ac.uk

† federico.lombardo@uniroma3.it

I. INTRODUCTION

A sequence of real numbers is called random if its statistical properties can provide insight into what constitutes "typical" behavior of real data obtained from a random experiment [1]. In principle, large amounts of such numbers can be used to solve any problem having a probabilistic interpretation by means of statistical sampling techniques. Therefore, it is needless to assert their usefulness in many different kinds of applications for the past seventy years, as the availability of computers made such statistical methods very practical [2].

Up to date, several algorithms have been put forward to produce computer-generated sequences of numbers that closely resemble the samples of independent and identically distributed (iid) random variables [3]. However, in many areas of physics and engineering, it is required to simulate stochastic processes with prescribed dependence structures [4]. For the majority of applications, the stochastic models are based on the idea that a time series in which successive values are correlated can frequently be regarded as generated from a Gaussian white noise into a linear filter [5].

Such an approach, often called the convolution method, allows one to produce sequences of real numbers with any arbitrary mean and autocorrelation function, if it is mathematically feasible. Conversely, the problem of generating correlated binary sequences with specified mean is still lacking a general solution, despite being a key issue in a variety of applications such as signal processing [6], modeling rainfall intermittency [7], the study of two-phase random media [8], just to name a few. This difficulty depends on the discrete (dichotomous) nature of binary processes, which makes the convolution method developed for processes defined over a continuous state space inapplicable [9]. Several techniques have been proposed to solve this problem. However, most of the existing methods demand a serious restriction on the class of autocorrelation functions that can be effectively modeled [9–12]. This paper presents an alternative robust algorithm to generate binary sequences with specified mean and autocorrelation function. It exploits the duality between the target binary process and the parent continuous process of transition probabilities to restate the problem in a continuous state space, thus allowing the application of spectrum-based iterative amplitude adjusted Fourier transform (IAAFT) method [13–15] to simulate continuous processes as a building block of an algorithm for binary random processes.

In the following, we firstly recall the theoretical concepts supporting the link between

binary processes and the corresponding parent transition probabilities as well as the distributional properties of such probabilities. We therefore use these properties to derive a simulation algorithm of binary signals based on the generation of sequences of parent transition probabilities defined in the continuous state space $[0, 1]$. Finally we show the algorithm performance for random processes with power-law and exponentially decaying autocorrelation functions along with a real world application involving the simulation of rainfall intermittency.

II. METHODOLOGY

A. Properties of binary process and parent transition probabilities

The problem is to generate a correlated sequence of random numbers $\{x_j\}_{j\in\mathbb{N}}$, for simplicity $\{x\}$, taking values 1 and 0 with probability p and $(1-p)$, respectively. The underlying discrete-time stochastic process X_j with state space $\{0, 1\}$, where j $(= 0, 1, 2, ...)$ denotes discrete time, is specified in terms of its mean $\mu_X = \mathrm{E}[X_j] = p$ and autocovariance function (ACVF)

$$c_X(\tau) = \mathrm{E}\left[X_j X_{j+\tau}\right] - \mu_X^2 = \sigma_X^2 \rho_X(\tau), \tag{1}$$

where $\mathrm{E}[\cdot]$ denotes expectation (ensemble average), τ is the time lag, $\sigma_X^2 = p(1-p)$ and $\rho_X(\tau)$ are the variance and autocorrelation function (ACF) of X_j, respectively. Denoting a generic sequence of uncorrelated values as $\{\varepsilon\}$, for continuous processes, the well-known convolution method allows the simulation of correlated sequences from $\{\varepsilon\}$ [16].

However, the direct application of the convolution method to $\{\varepsilon\}$ cannot produce binary random numbers, because the output of a linear filter is a non-binary sequence even if the input is binary [9]. To overcome this problem, we consider the conditional probability $Q_j = \Pr[X_j = 1 | \{\varepsilon\}]$ of occurring 1 at the jth place in the target sequence $\{x\}$, given the input $\{\varepsilon\}$. The sequence of conditional probabilities $\{q\}$ is a sample of the discrete-time and continuous-state stochastic process Q_j defined on the interval $[0, 1]$. Given $\{q\}$, the jth element of the sequence $\{x\}$ is generated by comparing each value q_j to a random number u_j sampled from a standard uniform distribution defined in $[0, 1]$, such as

$$x_j = \begin{cases} 1 & \text{if } \ u_j < q_j \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

Therefore, a correlated binary sequence can easily be generated if a suitable sequence of conditional probabilities $\{q\}$ is available. The algorithm proposed in this study exploits the correspondence between these two sequences by focusing on the simulation of the continuous process Q_j rather than the discrete one X_j, as the convolution method can be directly applied to synthesize the former. This implies the preliminary identification of the properties required for $\{q\}$ in order to be a sequence of conditional probabilities related to the target sequence $\{x\}$. In particular, it can be shown that the ACVF of the process Q_j equals that of the process X_j [9, 12]. Without loss of generality, it is convenient to assume that the two processes also have the same mean, $\mu_X = \mu_Q$, and variance, $\sigma_X^2 = \sigma_Q^2$.

In summary, we need to simulate a discrete-time stochastic process Q_j with prescribed ACF $\rho_Q(\tau) = \rho_X(\tau)$, and a continuous marginal distribution supported on the interval $[0, 1]$ with mean $\mu_Q = p$ and variance $\sigma_Q^2 = p(1-p)$. A suitable model for such marginal properties is the probability density function (pdf) of the beta distribution [17]

$$g(q; \alpha, \beta) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \tag{3}$$

where $q \in [0, 1]$, $\alpha > 0$ and $\beta > 0$ are two shape parameters, and the beta function $\mathrm{B}(\alpha, \beta)$ is a normalizing constant of the form

$$\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t. \tag{4}$$

We can express the parameters α and β of the distribution in terms of its mean μ_Q and variance σ_Q^2 as follows

$$\begin{cases} \alpha = \mu_Q \left(\dfrac{\mu_Q(1-\mu_Q)}{\sigma_Q^2} - 1 \right) \\ \beta = (1-\mu_Q) \left(\dfrac{\mu_Q(1-\mu_Q)}{\sigma_Q^2} - 1 \right) \end{cases}. \tag{5}$$

Substituting $\mu_Q = p$ and $\sigma_Q^2 = p(1-p)$ into eq. 5, we find that both parameters $\alpha = \beta = 0$ do not satisfy the condition of strict positiveness and the function $\mathrm{B}(\alpha, \beta)$ is undefined. When both parameters are less than one $(\alpha, \beta < 1)$, the beta distribution is U-shaped and

approaches a two-point Bernoulli distribution with equal probability masses $1/2$ at each end of the domain $[0,1]$ as $\alpha, \beta \to 0$ [18]. We seek new values of the parameters for which the continuous beta distribution of Q_j can mimic the discrete Bernoulli distribution of X_j with probability masses p and $(1-p)$ at 1 and 0. Therefore, we consider an arbitrarily small $\xi > 0$ such as

$$\frac{\mu_Q(1-\mu_Q)}{\sigma_Q^2} - 1 = \xi \to 0. \tag{6}$$

Substituting eq. 6 into eq. 5, we obtain the new parameter set as

$$\begin{cases} \alpha = \mu_Q \xi \\ \beta = (1-\mu_Q)\xi \end{cases}, \tag{7}$$

implying that $\mu_Q = \mu_X = p$ and $\sigma_Q^2 < p(1-p) = \sigma_X^2$. An extensive numerical investigation showed that $\xi = 0.05$ guarantees U-shaped beta distribution with $\sigma_Q^2 \approx p(1-p)$ and optimal convergence rate for the simulation methodology described in the following.

B. Simulation algorithm

Before describing each step of the proposed algorithm in detail, it is worth stressing that the theoretical considerations discussed in the previous section result in a conceptually very simple simulation procedure. It consists of generating a sequence of conditional probabilities $\{q\}$ following the beta distribution with parameters as in eq. 7 and the same ACVF as the target process X_j, and then applying the selection rule in eq. 2 to each value q_j in order to transform the sequence of transition probabilities into binary random numbers. The sequence $\{q\}$ is simulated by setting up IAAFT so that the spectral amplitudes corresponding to a correlated signal with Gaussian marginals are combined with the intensity of an uncorrelated sequence of values drawn from the required beta distribution. This way, IAAFT yields a signal with required ACF and marginal distribution. It is worth noting that though Q_j is defined as a stochastic process with random variables being conditional probabilities, the finite sequence $\{q\}$ is generated by reordering an uncorrelated time series $\{\varepsilon\}$, and an explicit computation of conditional probabilities is not required. Moreover, the modular structure of the algorithm allows one to use not only IAAFT but also other methods devised to simulate continuous processes with given marginal distribution and ACF, such as

the autoregressive-to-anything (ARTA) process generator [19] or the statically transformed autoregressive process (STAP) generator [20]. This further highlights the flexibility and generality of the proposed approach.

In more detail, the algorithm has the following steps:

1. Begin by using the convolution method to generate a sequence of n Gaussian random numbers $\left\{y_j\right\}_{j=0}^{n-1}$ with the desired ACF (e.g., power-law decay).

2. Store (i) the squared amplitudes of its Fourier transform, $S_k^2 = |\sum_{j=0}^{n-1} y_j \exp(i2\pi kj/n)|^2$, (ii) the sorted sequence $\left\{y_{(j)}\right\}_{j=0}^{n-1}$ where $y_{(j)}$ is the jth-smallest value of $\left\{y_j\right\}_{j=0}^{n-1}$, and (iii) a list of values $\left\{\varepsilon_j\right\}_{j=0}^{n-1}$ randomly drawn from the beta distribution with parameters in eq. 7.

3. Start the iteration procedure by reordering $\left\{\varepsilon_j\right\}_{j=0}^{n-1}$ to have the same rank structure as $\left\{y_j\right\}_{j=0}^{n-1}$, call the resulting sequence $\left\{q_j'^{(0)}\right\}_{j=0}^{n-1}$. Note that the two sequences share the same rank correlation, but not the desired linear ACF (or power spectrum). Each iteration m $(= 0, 1, 2, ...)$ consists of two consecutive steps:

   (a) First, the power spectrum of $\left\{q_j'^{(m)}\right\}_{j=0}^{n-1}$ is adjusted to that of $\left\{y_j\right\}_{j=0}^{n-1}$ by taking the Fourier transform of $\left\{q_j'^{(m)}\right\}_{j=0}^{n-1}$, replacing its squared amplitudes $\left\{S_k^{2,(m)}\right\}$ by $\left\{S_k^2\right\}$, and then transforming back. The phases are kept unaltered.

   (b) After this first step, $\left\{q_j'^{(m)}\right\}_{j=0}^{n-1}$ has the desired power spectrum but its marginal distribution has been modified. Therefore, in the second iterative step, the marginal distribution is adjusted by ordering $\left\{\varepsilon_j\right\}_{j=0}^{n-1}$ to have the same ranking as $\left\{q_j'^{(m)}\right\}_{j=0}^{n-1}$.

4. Since the power spectrum of the resulting sequence $\left\{q_j'^{(m+1)}\right\}_{j=0}^{n-1}$ is again modified, both iterative steps are repeated until a convergence threshold is achieved (here, mean absolute error equal to $5 \cdot 10^{-6}$ for S_k^2 is used in the numerical examples below).

5. Apply the selection rule in eq. 2 to transform the sequence of conditional probabilities into a binary sequence.

The algorithm stops with the exact matching of the beta marginal distribution, defined in $[0, 1]$, to properly apply the rule in eq. 2. This corresponds to "IAAFT-1" setup suggested by Kugiumtzis [14] when detailed properties of the amplitude distribution should be preserved.

The above algorithm differs from others previously proposed in the literature as it focuses on the simulation of the parent continuous-state process Q_j rather than on the target dichotomous process X_j, thus allowing the use of classical techniques such as IAAFT as the core of the simulation procedure. In this respect, it is worth noting that IAAFT (covering steps 1-4) has a very specific setup in this context. In fact, it is usually applied to obtain surrogate data preserving both marginal distribution and autocorrelation of a reference signal. Instead, the aim here is to combine the target marginal distribution in eq. 3 of a white noise $\{\varepsilon\}$ with the target power spectrum (ACF) of a Gaussian process Y_j, whose realization $\{y\}$ is generated by the classical convolution method. The resulting sequence $\{q'\}$ resembles the desired conditional probabilities $\{q\}$, which are used to generate binary random numbers $\{x\}$ with prescribed mean and ACF. Mathematically, there is a one-to-one correspondence between the marginal distributions of the processes Q_j and Y_j such as

$$q_j = h(y_j) = G^{-1}(\Phi(y_j); \alpha, \beta) \tag{8}$$

where G and Φ are the beta and Gaussian cumulative distribution functions of Q_j and Y_j, respectively. It should be stressed that the continuity of the state space of Q_j – its marginal pdf in eq. 3 is defined on the continuum between 0 and 1 – allows eq. 8 to hold true for any arbitrary dependence structure of Y_j. On the other hand, for discrete-type random variables such a relationship is not available [9]. This highlights the importance of moving from the binary process X_j to the continuous one Q_j for simulation purposes.

III. NUMERICAL EXAMPLES

A. Simulation of signals with exponentially and power-law decaying ACF

The performance of the proposed algorithm is tested by generating binary sequences with ACF corresponding to two stationary processes of paramount importance in several applications, i.e. the Hurst-Kolomogorov (HK) and the Markov process. The former, also known as fractional Gaussian noise, is characterized by the following ACF

$$\rho_X(\tau) = \frac{1}{2}(|\tau+1|^{2H} - 2|\tau|^{2H} + |\tau-1|^{2H}), \tag{9}$$

which exhibits a power-law decay $\rho_X(\tau) \propto |\tau|^{2H-2}$ and the corresponding power spectrum

takes the form $S_X(f) \propto f^{1-2H}$, with f the frequency and $H \in (0,1)$ the Hurst coefficient. This is analogous to the so-called pure power-law noises or $1/f$ noises. For $0.5 < H < 1$ the process is positively correlated and exhibits long-range dependence, while it reduces to white noise for $H = 0.5$. As a second example, we consider a process with short-range Markovian dependence, which is characterized by exponentially decaying ACF of the form

$$\rho_X(\tau) = \exp(-\gamma|\tau|) = \rho_1^{|\tau|}, \tag{10}$$

where $1/\gamma$ is the correlation radius and $\rho_1 = \exp(-\gamma)$ is the lag-one autocorrelation coefficient. Simulations are performed for $H \in \{0.7, 0.8, 0.9\}$, $\rho_1 \in \{0.5, 0.8, 0.95\}$, $p \in \{0.01, 0.1\}$, and sample size 2^{20}. The values of p may mimic the rate of occurrence of rare events such as storms, floods, earthquakes and other geophysical hazards. The large sample size was chosen to highlight whether the simulated samples approach the theoretical behavior as expected when the size tends to very large values. Results for HK and Markov processes are illustrated in figs. 1 and 2, respectively. Left panels compare the theoretical ACFs with the empirical counterpart with the set of parameters mentioned above. The agreement between theoretical and simulated ACFs and the lack of scattering of the empirical ACF values denote the effectiveness of the proposed approach to simulate binary signals with power-law and exponential decay as well as prescribed mean and variance.

For each model, panels of figs. 1 and 2 in the right side show some examples of synthetic sequences of equal length. In all cases, values 0 and 1 tend to cluster more and more as the degree or extent of autocorrelation increase. This clustering behavior is in agreement with previous theoretical and empirical findings resulting from the study of extreme values of simulated and observed processes taking real values [21–29]. In this respect, we stress again that the simulated sequences preserve on average not only the desired ACF but also the desired rate of 0 and 1 of the underlying process, thus allowing the study of the sampling variability for finite size sequences. This is of paramount importance to mimic and study for instance the occurrence of extreme geophysical phenomena and the related sampling uncertainty in order to provide more reliable risk assessment.

It should be noted that the variability of the empirical ACFs for the sample size 2^{20} is negligible for ACF values higher than ≈ 0.02. For $p = 0.01$, some lack of accuracy emerges especially for low ACF values and weaker correlation structure corresponding to $H = 0.7$ and $\rho_1 = 0.5$. However, some fluctuations are expected in this case because of the sharpness of

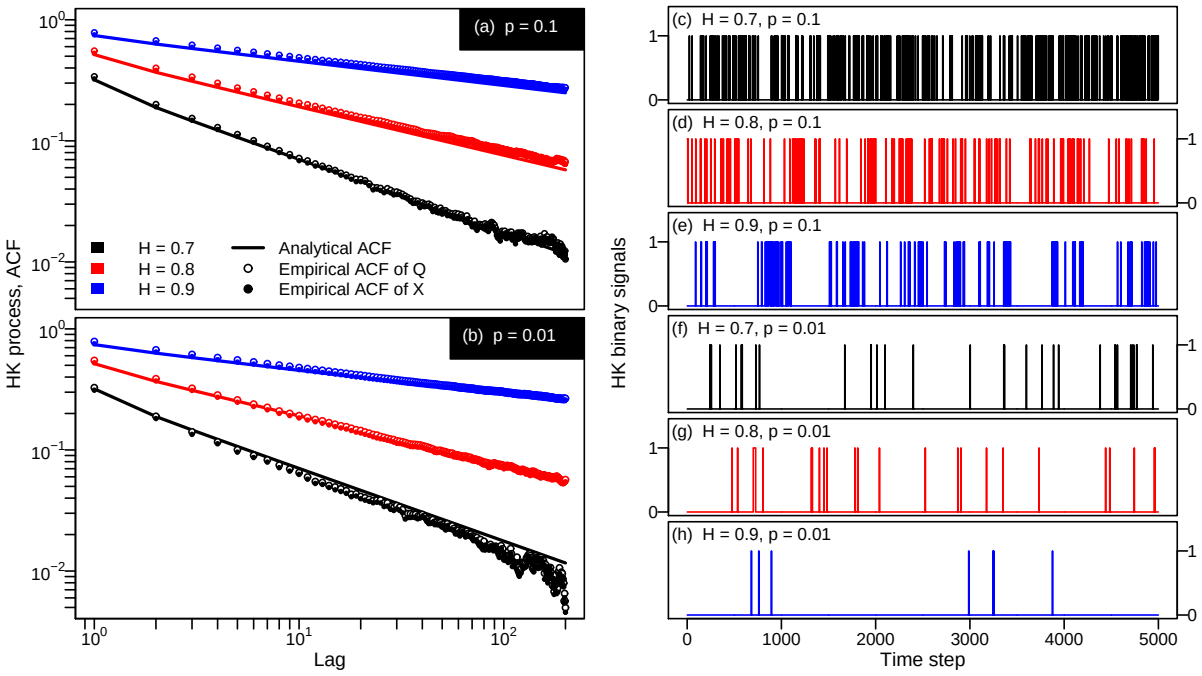

FIG. 1. ACFs (a-b) and sample signals (c-h) corresponding to HK processes with three different values of the characteristic parameter $H \in \{0.7, 0.8, 0.9\}$, and $p \in \{0.01, 0.1\}$. Panels (a-b) show the analytical ACFs (—) along with the empirical ACFs of the simulated sequences of transition probabilities $\{q\}$ (○) and binary signal $\{x\}$ (•) for each of the parameter values reported in the legends. Panels (c-e) and (f-h) depict synthetic sequences corresponding to ACFs reported in (a) and (b), respectively. The simulated sequences exhibit an increasing clustering effect related to the increasing strength of the autocorrelation (viz. parameter values).

the beta distribution, whose probability mass is concentrated very close to the boundaries of its domain (zero and one), as well as finite-size effects affecting the empirical ACF of sequences with very low rate of occurrence.

We further explored finite size effects by simulating binary sequences of size 10^3 to 10^4 by steps of 10^3 for $H \in \{0.7, 0.8, 0.9\}$, $\rho_1 \in \{0.5, 0.8, 0.95\}$, $p \in \{0.01, 0.05, 0.1\}$, and then assessing the variability and bias of p and ACF, as well as the behavior of the number of IAAFT iterations required to reach convergence. For each combination of parameters, 100 time series were generated. For HK process, fig. 3 shows that the simulated sequences exhibit unbiased p values for each combination of the parameters with variability decreasing as the sample size increases. Similar results (not shown) hold for Markov process and are expected as the algorithm reproduces almost exactly the Bernoulli marginal distribution

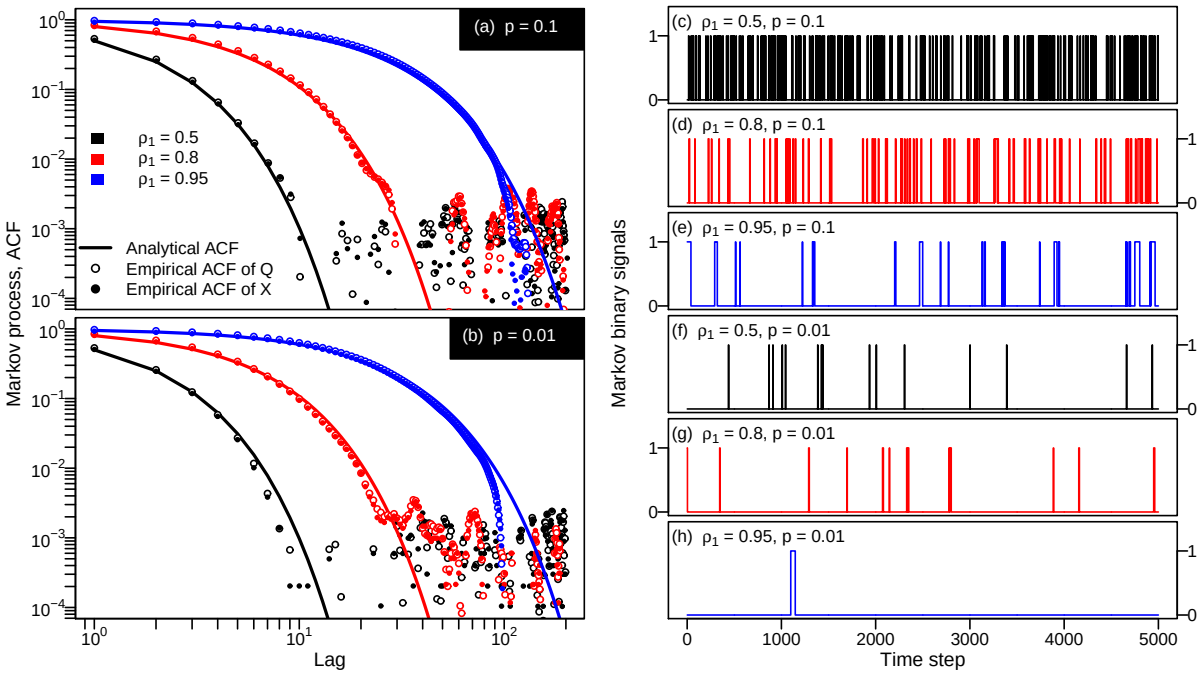

FIG. 2. As fig. 1 but for Markov process with parameter $\rho_1 \in \{0.5, 0.8, 0.95\}$; the same caption and interpretation apply. Empirical ACF values exhibit strong fluctuations and lose their agreement with the theoretical curves for $\rho_X(\tau) \approx 0.005$ because of finite size effects.

with rate of occurrence p.

For HK processes, fig. 4 shows the mean error between the theoretical ACF and the empirical ACFs computed on lags from one to 11 in order to emphasize the contribution of the larger ACF terms. In this case, we have a residual negative bias decreasing as p increases, and variability increasing with H. The ACF bias converges to zero as the sample size increases. This convergence is also evident for Markov processes (fig. 5). It should be noted that such a bias is not a limit of our algorithm, but it is due to the estimation of the autocorrelation function from data. In fact, this is characterized by negative bias, which may be very high when the process exhibits long-term persistence (i.e., HK process) [30, 31]. This is particularly the case with binary processes that are characterized by sequences of thousands of zero values for low p values (see bottom panels of figs. 1 and 2). This might easily prevent the recognition of the actual correlation structure if the sample size is not large enough. This behavior further justifies the choice of sequences of size 2^{20} to evaluate the actual agreement between the properties of the simulated signals and those of the underlying processes.

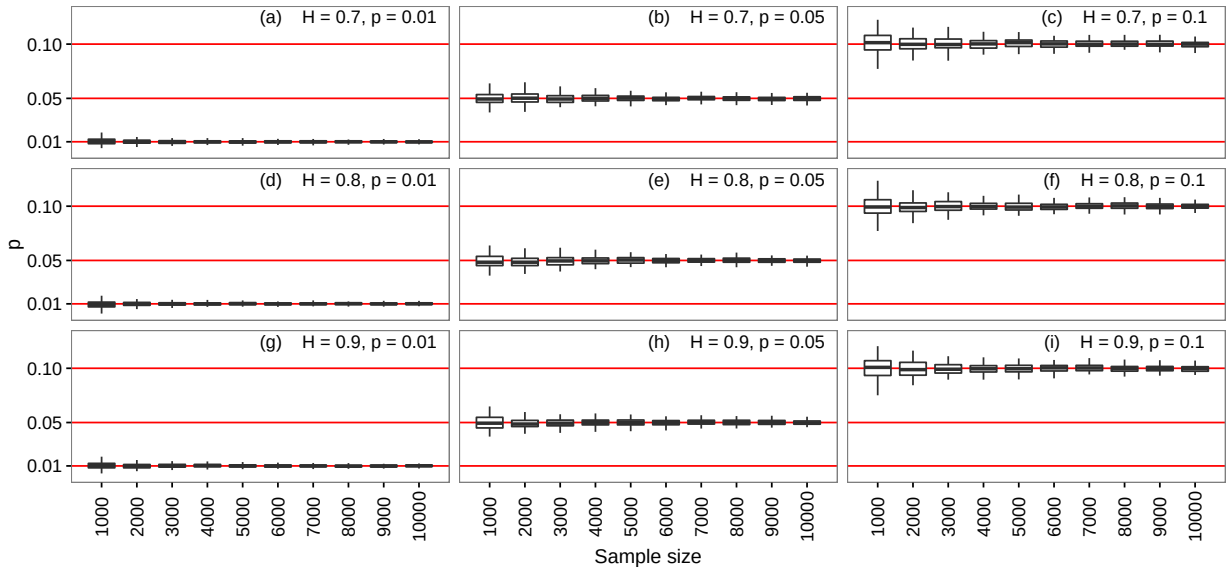

FIG. 3. Box plots showing the variability of p computed on sequences of size 10^3 to 10^4 by steps of 10^3 for HK processes with $H \in \{0.7, 0.8, 0.9\}$, and $p \in \{0.01, 0.05, 0.1\}$. Finite size does not introduce any bias, while the variance of p decreases as the sample size increases.

For HK process, fig. 6 shows that the expected number of IAAFT iterations required to achieve convergence increases with the sample size and p, while it is almost independent of H. Similar results hold for Markov process (not shown). Of course, the overall number of iterations globally increases or decreases based on the tolerance of the convergence criterion used in the IAAFT algorithm.

B. Simulation of rainfall intermittency

As mentioned in the introduction, binary sequences are very common in physics and geophysics as they naturally rise when one focuses on the occurrence/non-occurrence or presence/absence of a given event and/or characteristic. A matter of common experience is the rainfall intermittency, i.e. the alternation of wet and dry periods. The dependence structure of the rainfall occurrence process appears to be non-Markovian [7], and the reproduction of the observed rate of occurence, p, is of paramount importance in hydrological engineering. Therefore, a general simulation method reproducing the moments of the occurrence process up to the second order is of practical interest. For the sake of illustration, we consider a rainfall time series recorded at Casigliano (central Italy) at 30-minute temporal

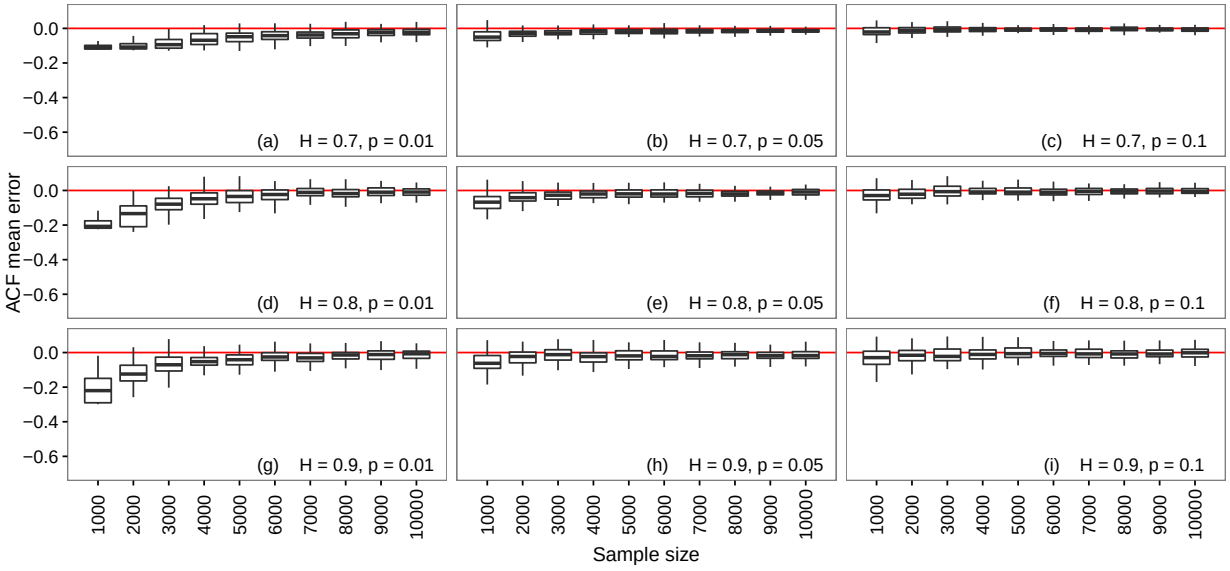

FIG. 4. Box plots summarizing the variability of ACF mean error computed on sequences of size 10^3 to 10^4 by steps of 10^3 for HK processes with $H \in \{0.7, 0.8, 0.9\}$, and $p \in \{0.01, 0.05, 0.1\}$. Finite size introduces some bias that tends to decrease as the sample size increases. Notice that such a bias is not a limit of the proposed algorithm, but it is due to the estimation of the autocorrelation function from finite size sequences.

resolution from 1995 to 2001. As these data belong to a wider data set of 35 time series previously studied [32, 33], we refer the reader to the literature for further details. As the rainfall exhibits seasonal fluctuations, in order to have a homogenous sequence, we focused on October data (similar results can be obtained for the other months). Figure 7(a-b) shows the time series of rainfall occurrence and rainfall depth. The estimated rate of occurrence is $\hat{p} = 0.047$. Empirical ACF was computed on the merged sample comprising October records for the years 1995-2001, taking care of removing the cross-products of lagged observations, x_j and $x_{j+\tau}$, not belonging to the same year [34]. For the sake of comparison, the occurrence process was modeled by three different dependence structures; namely, HK (eq. 9) with $\hat{H} = 0.84$ estimated by the least squares based on variance (LSV) method [35], Markov (eq. 10) with $\hat{\rho}_1 = 0.69$, and purely random Bernoulli ($\rho_X(\tau) = 0$). Results are shown in figs. 7(c-j). The example time series in figs. 7(c-e) show that both Bernoulli and Markov processes cannot mimic the typical clustering behavior of rainfall occurrence, which is in turn well reproduced by HK. Figures 7(g-i) compare the empirical ACF of the observed process with the average ACF computed from the ACFs of 100 simulated sequences with the same

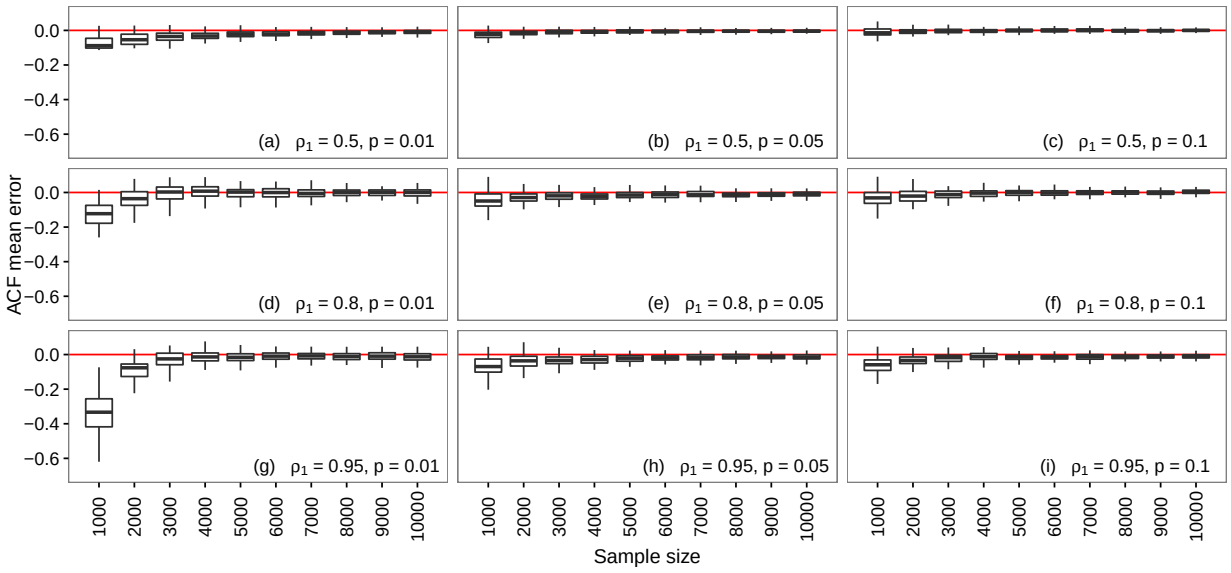

FIG. 5. As fig. 4 but for Markov processes with $\rho_1 \in \{0.5, 0.8, 0.95\}$; the same caption and interpretation apply. Notice that the convergence to theoretical ACF is faster than in the case of HK process (fig. 4), as the estimation bias of the second order moments due to finite size effects is larger for processes characterized by long-term persistence.

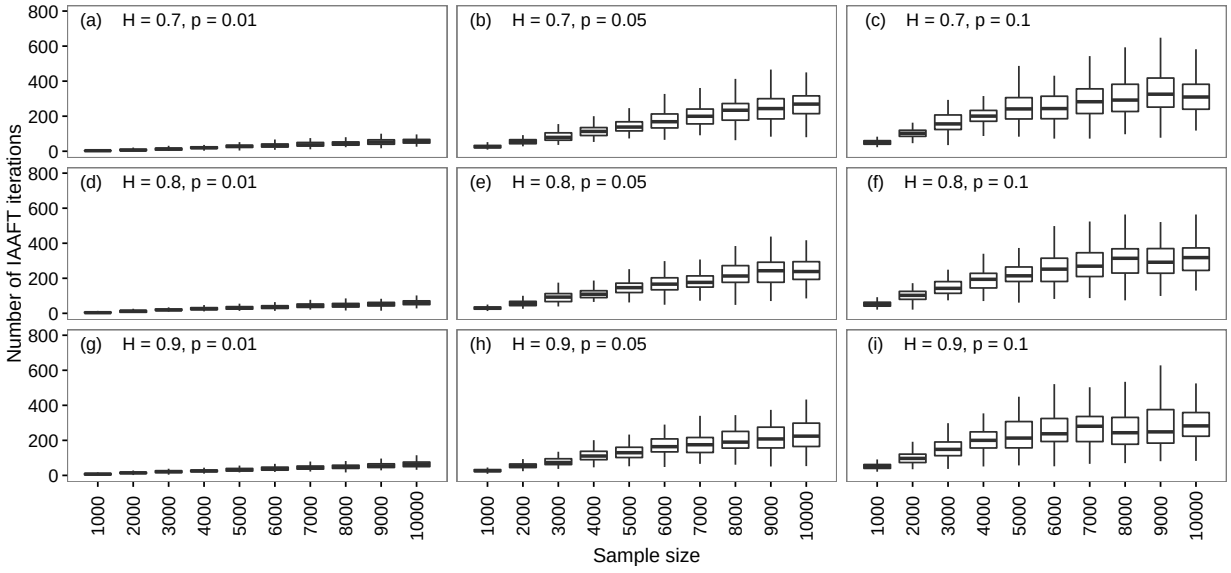

FIG. 6. Box plots showing the sampling variability of the number of IAAFT iterations required to achieve convergence for binary sequences with HK ACF, $H \in \{0.7, 0.8, 0.9\}$, and $p \in \{0.01, 0.05, 0.1\}$. The expected number of IAAFT iterations increases with the sample size and p, while it is almost independent of H. Similar results hold for Markov process (not shown).

size of the observed time series. The 90% confidence limits of the ACF highlight that the observed ACF is compatible with HK once the sampling uncertainty is taken into account, whereas pure randomness and Markovian dependence are not well suited for the data at hand. Moreover, the larger width of HK confidence bands denotes larger uncertainty, which reflects the larger variability of such a type of strongly persistent processes. This further confirms the importance of modeling and simulating binary signals with prescribed mean and autocorrelation.

Describing rainfall occurrence by theoretical processes allow for the set up of parametric models that can be used for prediction owing to their explanatory power, for sensitivity analysis (by varying model parameters), or as modules for more general frameworks devised for simulating the entire rainfall process (occurrence and intensity). However, for exploratory purposes (e.g., nonlinearity testing [13–15, 20]) it can be useful to have so-called surrogate data that preserve some key properties of the observed signals. We have therefore tested the capability of the proposed methodology to generate surrogate sequences preserving on average the observed $\hat{p}$ and empirical autocorrelation function. To this aim, the input Gaussian sequences required by the simulation algorithm are generated by drawing iid sequences from a standard Gaussian distribution and then introducing the empirical correlation using the Cholesky decomposition of the empirical correlation matrix. Figure 7(f) shows one of 100 surrogate series, while fig. 7(j) shows the empirical ACF along with the average simulation and the 90% confidence bands. These diagrams confirm that the proposed method can easily generate accurate surrogate series preserving on average the observed ACF with limited variability around the expected pattern. Even though further investigation of these aspects is required, these results confirm the flexibility, generality, and potentialities of the algorithm introduced in this study.

IV. CONCLUSIONS

As highlighted in Ref. [12], the discrete nature of binary signals can introduce theoretical constraints in classical simulation methods, thus limiting the generation of binary signals with prescribed autocorrelation. This apparently prevents the use of general algorithms, which are instead available for continuously distributed random sequences. The proposed method overcomes this limitation focusing on the parent continuous process of transition

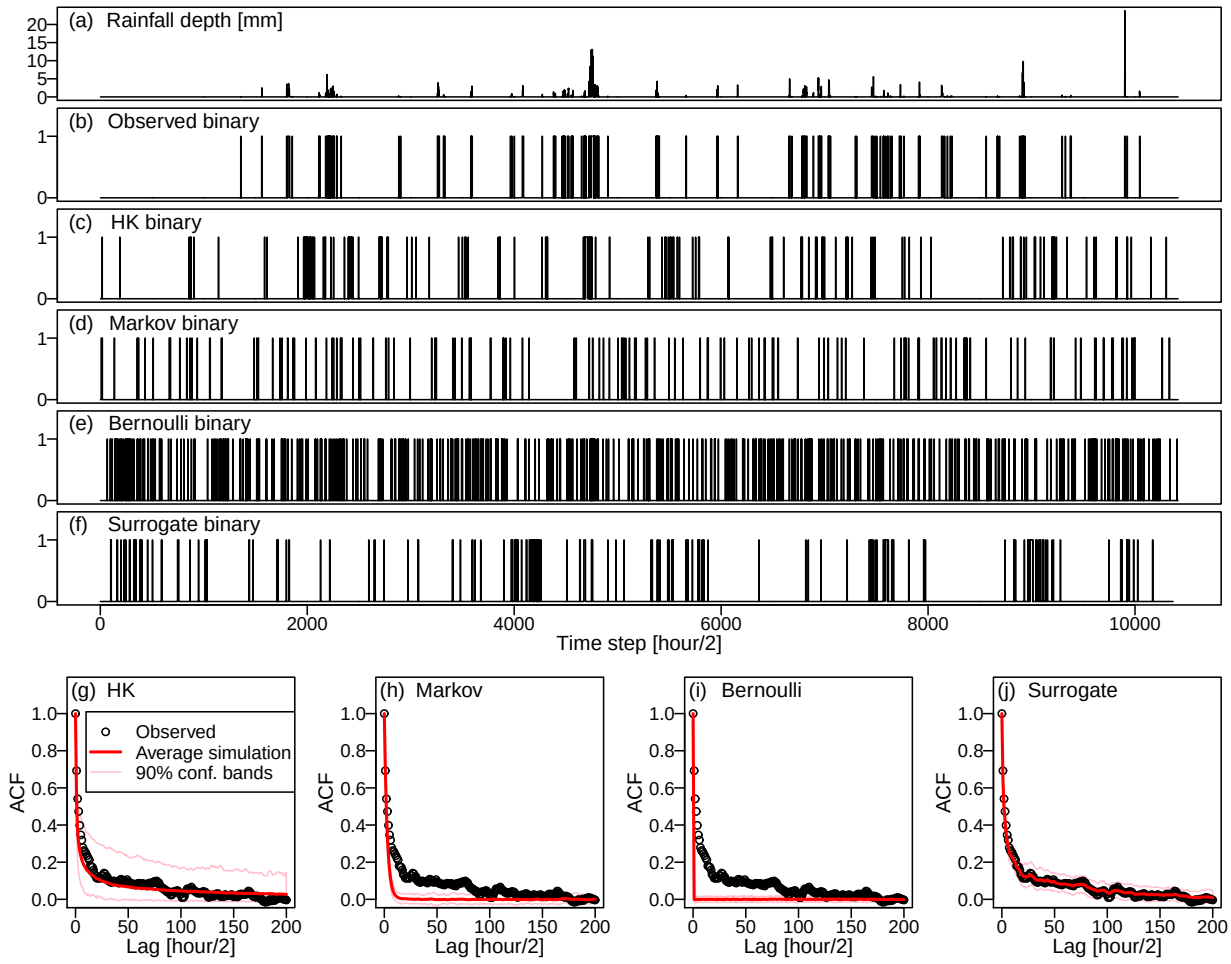

FIG. 7. (a) Time series of rainfall depth [mm] obtained by merging October data recorded at Casigliano (central Italy) at 30-minute temporal resolution from 1995 to 2001. (b) Observed occurrence process corresponding to the rainfall records in panel (a). (c-f) Typical synthetic time series, of equal length, generated by the proposed algorithm respectively with HK and Markov dependence structures, Bernoulli pure randomness, and empirical ACF (surrogate sequence). Comparison with the observed occurrence process in panel (b) shows that HK and surrogate can reproduce the typical clustering behavior (also known as over-dispersion) of rainfall events, whereas Markov and Bernoulli occurrences cover the time axis more homogeneously (indicating so-called equi- or under-dispersion). (g-j) Comparison of the ACF of the observed occurrence process and the mean ACF obtained by averaging the ACFs of 100 synthetic signals with the same size of the observed sequence. The 90% confidence bands are also reported. Panels (g-i) show that the HK average ACF fits very well the observed ACF, and sampling fluctuations fall within the confidence bands; on the other hand, Markov and Bernoulli dependence structures clearly underestimate the observed ACF. The average ACF of surrogate series in panel (j) closely follow the empirical ACF with limited fluctuations as expected according to the definition of surrogate.

probabilities rather than on the target binary process. Since the parent process is characterized by a continuous beta marginal distribution and a given correlation structure, the simulation problem is moved back from discrete to continuous state space, thus allowing for use of classical convolution techniques and the corresponding freedom in terms of desired correlation structure. Once a sequence of correlated beta distributed random variables is generated, the corresponding binary sequence results from a simple acceptance/rejection criterion. As compared with simpler methods, the proposed approach allows one to specify and control not only the correlation structure but also mean and variance of the binary signal. This is a key aspect for practical applications involving for instance anthropogenic and natural hazards, such as rainfall events analyzed in this study. These extreme events exhibit a low rate of occurrence and are usually grouped into clusters. Reliable risk analyses should therefore require modeling both aspects.

ACKNOWLEDGMENTS

Francesco Serinaldi acknowledges support from Willis Research Network. Federico Lombardo is grateful to Silvia Ghinaglia for her continuous support and fruitful discussions. Remarks from two anonymous reviewers are gratefully acknowledged. The analyses were performed in `R` [36].

[1] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. (McGraw Hill, New York, 1991).

[2] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev, Wiley Interdisciplinary Reviews: Computational Statistics **6**, 386 (2014).

[3] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge, 2007).

[4] N. J. Kasdin, P. IEEE **83**, 802 (1995).

[5] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*, 5th ed. (John Wiley & Sons, Hoboken, New Jersey, 2015).

[6] S. R. Dey and A. V. Oppenheim, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 3 (2007) pp. 1493–1496.

[7] D. Koutsoyiannis, Water Resour. Res. **42**, W01401 (2006).

[8] P.-S. Koutsourelakis and G. Deodatis, J. Eng. Mech.-ASCE **131**, 397 (2005).

[9] F. M. Izrailev, A. A. Krokhin, N. M. Makarov, and O. V. Usatenko, Phys. Rev. E **76**, 027701 (2007).

[10] P. Boufounos, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 3 (2007) pp. 981–984.

[11] C. R. Rojas, J. S. Welsh, and G. C. Goodwin, in *2007 American Control Conference* (2007) pp. 122–127.

[12] O. V. Usatenko, S. S. Melnik, S. S. Apostolov, N. M. Makarov, and A. A. Krokhin, Phys. Rev. E **90**, 053305 (2014).

[13] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[14] D. Kugiumtzis, Phys. Rev. E **60**, 2808 (1999).

[15] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).

[16] D. Koutsoyiannis, Water Resour. Res. **36**, 1519 (2000).

[17] A. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications* (Taylor& Francis, 2004).

[18] M. Zhu and A. Y. Lu, J. Stat. Educ. **12**, 1 (2004).

[19] M. C. Cario and B. L. Nelson, Oper. Res. Lett. **19**, 51 (1996).

[20] D. Kugiumtzis, Phys. Rev. E **66**, 025201 (2002).

[21] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin, Phys. Rev. Lett. **94**, 048701 (2005).

[22] J. F. Eichner, J. W. Kantelhardt, A. Bunde, and S. Havlin, Phys. Rev. E **73**, 016130 (2006).

[23] M. I. Bogachev, J. F. Eichner, and A. Bunde, Phys. Rev. Lett. **99**, 240601 (2007).

[24] J. F. Eichner, J. W. Kantelhardt, A. Bunde, and S. Havlin, Phys. Rev. E **75**, 011128 (2007).

[25] M. I. Bogachev, J. F. Eichner, and A. Bunde, Eur. Phys. J.-Spec. Top. **161**, 181 (2008).

[26] J. F. Eichner, J. W. Kantelhardt, A. Bunde, and S. Havlin, in *In Extremis*, edited by J. Kropp and H.-J. Schellnhuber (Springer Berlin Heidelberg, 2011) pp. 2–43.

[27] M. I. Bogachev and A. Bunde, Europhys. Lett. **97**, 48011 (2012).

[28] E. Volpi, A. Fiori, S. Grimaldi, F. Lombardo, and D. Koutsoyiannis, Water Resour. Res. **51**, 8570 (2015).

[29] F. Serinaldi and C. G. Kilsby, Water **8**, 152 (2016).

[30] D. Koutsoyiannis, Hydrolog. Sci. J. **48**, 3 (2003).

[31] D. Koutsoyiannis and A. Montanari, Water Resour. Res. **43**, W05429 (2007).

[32] F. Serinaldi, Stoch. Env. Res. Risk A. **22**, 671 (2008).

[33] F. Serinaldi, Stoch. Env. Res. Risk A. **23**, 677 (2009).

[34] S.-M. Papalexiou, D. Koutsoyiannis, and A. Montanari, J. Hydrol. **411**, 279 (2011).

[35] H. Tyralis and D. Koutsoyiannis, Stoch. Env. Res. Risk A. **25**, 21 (2011).

[36] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2015), ISBN 3-900051-07-0.