Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions

Ioannis Tsoukalas, Christos Makropoulos and Demetris Koutsoyiannis

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

Key Points:

- Simulation of processes with arbitrary marginal distributions
- Simulation of short and long-range dependent processes
- Simulation of univariate and multivariate processes

Keywords:

multivariate stochastic simulation, short and long-range dependence, Hurst phenomenon, symmetric moving average to anything, Nataf joint distribution model, arbitrary marginal distributions

Index terms: 1869, 3265, 3270, 1872, 1880

Supporting information:

https://agupubs.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1029%2F2017WR022462&file=wrcr23687-sup-0001-2017WR022462_S01.pdf

Citation:

Tsoukalas, I., Makropoulos, C., & Koutsoyiannis, D. (2018). Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions. *Water Resources Research*. https://doi.org/10.1029/2017WR022462

Journal: Water Resources Research
Received 21 DEC 2017
Accepted 29 OCT 2018
Accepted article online 05 NOV 2018

Abstract

Hydrometeorological processes are typically characterized by temporal dependence, short- or long-range (e.g., Hurst behavior), as well as by non-Gaussian distributions (especially at fine time scales). The generation of long synthetic time series that resemble the marginal and joint properties of the observed ones is a prerequisite in many uncertainty-related hydrological studies, since they can be used as inputs and hence allow the propagation of natural variability and uncertainty to the typically deterministic water-system models. For this reason, it has been for years one of the main research topics in the field of stochastic hydrology. This work presents a novel model for synthetic time series generation, termed Symmetric Moving Average (neaRly) To Anything (SMARTA), that holds out the promise of simulating stationary univariate and multivariate processes with any-range dependence and arbitrary marginal distributions, provided that the former is feasible and the latter have finite variance. This is accomplished by utilizing a mapping procedure in combination with the relationship that exists between the correlation coefficients of an auxiliary Gaussian process and a non-Gaussian one, formalized through the Nataf's joint distribution model. The generality of SMARTA is stressed through two hypothetical simulation studies (univariate and multivariate), characterized by different dependencies and distributions. Furthermore, we demonstrate the practical aspects of the proposed model through two real-world cases, one that concerns the generation of annual non-Gaussian streamflow time series at four stations, and another that involves the synthesis of intermittent, non-Gaussian, daily rainfall series at a single location.

1 Introduction

Hydrometeorological time series (i.e., sequences of observations ordered in time) can be considered the cornerstone of *any* water-related engineering study, although, such data are in scarcity and often the available records don't have sufficient length for the task at hand (e.g., reliability and risk-related studies). A historical record of such observations will rarely if ever repeat in the future, which is the simplest manifestation of the high variability and uncertainty that is naturally inherited therein. In this vein, it can be argued that embracing stochasticity in hydrometeorological processes is a first step towards the development of uncertainty-aware methodologies for water systems. Stochastic simulation, and the synthesis of long hydrometeorological time series, which are used in place of historical ones, can provide a potential remedy to this situation. Synthetic time series are not predictions of future states, but rather constitute plausible realizations of the simulated process, that are, loosely speaking, statistically equivalent with the parent information (i.e., historical data). Driving the typically deterministic water-system simulation models with such realizations provides the means to assess their response in a probabilistic manner, under multiple, plausible scenarios. Nowadays, synthetic data are used in a variety of studies, among them, the optimal planning and management of reservoir systems (e.g., Celeste & Billib, 2009; Feng et al., 2017; Giuliani et al., 2014; Koutsoyiannis & Economou, 2003; Tsoukalas & Makropoulos, 2015a, 2015b), risk assessment of flood (e.g., Haberlandt et al., 2011; Paschalis et al., 2014; Qin & Lu, 2014; Wheater et al., 2005) and drought events (e.g., Herman et al., 2016), as well as water resources simulation under future climate conditions (e.g., Fatichi et al., 2011; Fowler et al., 2000; Kilsby et al., 2007; Nazemi et al., 2013). Thereby, the wide applicability of synthetic time series and stochastic simulation highlight the need for simulation schemes that can resemble the, intriguing and challenging to simulate, characteristics of hydrometeorological processes.

A typical characteristic encountered in such processes is auto-dependence (persistence), either short or long-range. The former, short-range dependence (SRD), has been extensively discussed in literature (e.g., Box et al., 2015) and implies an exponential autocorrelation structure that diminishes after few time lags. On the contrary, the second, long-range dependence (LRD), also known as long-term persistence (sometimes referred to as long-memory), implies an auto-dependence structure that strongly extends for large lags (see, Beran, 1992). The latter behavior is also related to the so-called Hurst phenomenon, also known as Joseph effect, fractional Gaussian noise (fGn), scaling in time or Hurst-Kolmogorov dynamics (HK; Koutsoyiannis, 2011; Koutsoyiannis & Montanari, 2007); see also the review work of Molz et al. (1997). Its discovery is usually credited to Hurst (1951), who while studying long records of streamflow and other data noticed that extreme events tend to exhibit a clustering behavior in terms of temporal occurrence. However, as pointed out by Koutsoyiannis (2011), it was Kolmogorov (1940) who first proposed its mathematical description. Eventually, after the seminal work of Hurst and the extensive documentation of Mandelbrot and Wallis (1969a, 1969b, 1969c) it is now acknowledged that LRD (and HK) processes are omnipresent in geophysics, hydrology, climate and other scientific disciplines (Beran, 1994; Koutsoyiannis, 2002; O'Connell et al., 2016). The latter publications provide further examples and details regarding the interpretation and identification of such behavior.

Regarding modelling and application of SRD or LRD in hydrological studies, the former type (SRD) has been systematically studied and employed in numerous cases for the simulation of a variety of hydrometeorological processes (Breinl et al., 2013; Brissette et al., 2007; Khalili et al., 2009; Matalas, 1967; Mehrotra et al., 2015; Mhanna & Bauwens, 2012; Srikanthan & McMahon, 2001; Srikanthan & Pegram, 2009; Thompson et al., 2007). On the other hand, it is well recognized that proper representation of LRD is of high importance, especially in reservoir-related studies, since their operation and regulation is performed in over-annual scale, where LRD is mostly encountered (Bras & Rodríguez-Iturbe, 1985; Iliopoulou et al., 2016; Koutsoyiannis, 2002). Other notable hydrology-related applications of LRD include the stochastic simulation of precipitation or streamflow at finer time-scales, from monthly and daily (e.g., Detzel & Mine, 2017; Maftei et al., 2016; Montanari et al., 1997, 2000) to 10-second interval (e.g., Lombardo et al., 2012; Papalexiou et al., 2011), as well as the generation of synthetic storm hyetographs (e.g., Koutsoyiannis & Foufoula-Georgiou, 1993).

In general, SRD can be easily captured with the so-called AutoRegressive Moving Average (ARMA) family of models, while we note that, even though such models have a long history and are still popular, today the literature offers other powerful and flexible options (cf. Koutsoyiannis, 2016). On the other hand, LRD, hence HK behavior, requires the use of alternative generation schemes (see, Bras & Rodríguez-Iturbe, 1985; O'Connell et al., 2016), such as fractional Gaussian noise models (Mandelbrot & Wallis, 1969a, 1969b, 1969c), fast fractional Gaussian noise (ffGn) models (Mandelbrot, 1971), broken line models (Ditlevsen, 1971; Mejia et al., 1972) and Fractional AutoRegressive Integrated Moving-Average (FARIMA) models (Granger & Joyeux, 1980; Hosking, 1984). In contrast to the abovementioned specialized simulation schemes, a notable exception, that can simulate any type of autocorrelation function of a process, is the Symmetric Moving Average (SMA) model of Koutsoyiannis (2000, 2002, 2016), coupled with theoretical autocorrelation (or autocovariance) structures. This flexibility is achieved by decoupling the parameter identification of the autocorrelation structure and the generation mechanism (i.e., the model).

In addition to temporal dependence, hydrometeorological variables are often characterized by non-Gaussian and skewed distribution functions (partially attributed to the often non-negative nature of such processes), especially in fine time scales (e.g., daily or finer), where intermittency is omnipresent. The need to account for non-Gaussian distributions was early recognized by many researchers (e.g., Klemeš & Borůvka, 1974; Matalas & Wallis, 1976; Matalas, 1967) and is currently remarked by the numerous large-scale statistical studies conducted at various time scales (e.g., Blum et al., 2017; Cavanaugh et al., 2015; Kroll & Vogel, 2002; McMahon et al., 2007; Papalexiou & Koutsoyiannis, 2013, 2016). Regarding stochastic hydrology and simulation through linear stochastic models, many efforts have been made towards that direction (i.e., simulating non-Gaussian processes) which can be broadly classified in three main categories (Tsoukalas et al., 2018a): a) Explicit methods that are able to generate data from specific marginal distributions (e.g., Klemeš & Borůvka, 1974; Lawrance & Lewis, 1981; Lombardo et al., 2012, 2017; Matalas, 1967) b) Implicit approaches, pioneered by Thomas and Fiering (1963), that treat skewness via employing non-Gaussian white noise (typically from Pearson type-III distribution) for the innovation term (Detzel & Mine, 2017; Efstratiadis et al., 2014; Koutsoyiannis, 1999, 2000; Lettenmaier & Burges, 1977; Matalas & Wallis, 1976, 1971; Matalas, 1967; Todini, 1980). c) Transformation-based approaches that employ appropriate functions (e.g., Box-Cox) in order to "*normalize*" the observed data; next simulate realizations using typical Gaussian stochastic models and finally "*de-normalize*" the generated data in order to attain the process of interest (e.g., Salas et al., 1980). However, as discussed in Tsoukalas et al. (2018a), most of these schemes exhibit a number of limitations that still remain unresolved. Particularly, approaches of category (a) are limited to a narrow type of autocorrelation functions and non-Gaussian distributions (e.g., Gamma or Log-Normal), while they are typically able to simulate only univariate processes. On the other hand, approaches of category (b) are prone to the generation of negative values, provide an approximation of the marginal distributions, while encounter difficulties when modelling highly skewed (univariate or multivariate) processes (Koutsoyiannis, 1999; Todini, 1980). It is noted thought, that some recent schemes are able to capture moments higher than skewness (e.g., kurtosis), by the inclusion of additional model parameters (Koutsoyiannis et al., 2018 and references therein). On top of these issues, only few schemes (e.g., SMA) are able to model a variety of temporal correlation structures, while it is also possible to establish bounded dependence patterns which are far from natural ones (Tsoukalas et al., 2018a, 2018b). Finally, regarding the schemes of category (c), they require the specification of a non-trivial normalization function (due to the inadequacy of simple transformations; such as, Box-Cox) that often entail several parameters (usually determined through optimization techniques). Further to this, even if the latter function is properly identified, it is acknowledged that they introduce bias in the simulated marginal and joint characteristics (Bras & Rodríguez-Iturbe, 1985; Salas et al., 1980 p. 73).

In this work, in an effort to simultaneously address these challenges and provide a flexible method for synthetic time series generation, we introduce a generic, yet simple and theoretically justified, explicit approach based on the simulation of univariate and multivariate stationary processes exhibiting any-range dependence and arbitrary marginal distributions. More specifically, the proposed method can explicitly model the autocorrelation structure and distribution of each individual process, provided that the former is feasible and the latter have finite variance, while simultaneously it can preserve the lag-0 cross-correlation structure. The main components of the method are, the SMA model of Koutsoyiannis (2000), a theoretical autocorrelation structure and the pivotal concept of Nataf's joint distribution model (NDM, Nataf, 1962). The key idea of our approach lies in employing an auxiliary Gaussian stochastic process, modelled using the SMA scheme, with such parameters that reproduce the target auto- (i.e., temporal; SRD or LRD) and lag-0 cross-correlation (i.e., spatial) coefficients of the process after its subsequent mapping to the actual domain via the target inverse cumulative density functions (ICDFs). It is remarked that instead of SMA, any other linear stochastic model (e.g., ARMA-type) could be employed in order to mathematically describe the auxiliary Gaussian process, yet, it is anticipated that the resulting simulation scheme will inherit its properties regarding the simulation of univariate and multivariate processes, e.g., if the auxiliary model is capable of simulating SRD structures, the established simulation scheme will also be SRD.

The latter rationale has also been employed within the scientific field of operations research and particularly by Cario and Nelson (1996), as well as, Biller and Nelson (2003) who proposed the AutoRegressive To Anything (ARTA) and the Vector AutoRegressive To

Anything (VARTA) methods respectively for the explicit simulation of stationary autoregressive (AR) processes with arbitrary marginal distributions.

It is remarked that (to the extent of our knowledge) despite their wide acceptance, the aforementioned approaches (and their variants) have been unknown to the hydrological community and have never been used for the simulation of hydrometeorological processes until very recently. Nonetheless, it seems that presently, Nataf-based approaches are gaining momentum. Particularly, using a similar rationale, Serinaldi and Lombardo (2017) introduced an approach for the synthesis of autocorrelated univariate binary processes, while, Papalexiou (2018) provided a comprehensive treatment on the topic using autoregressive models and used, for first time, mixed-type marginals enabling the modeling of intermittent processes like precipitation. Finally, Tsoukalas et al. (2017, 2018a), employed the notion of NDM and provided a generalization of the latter models (ARTA, VARTA), termed SPARTA (Stochastic Periodic AutoRegressive To Anything), for the simulation of univariate and multivariate cyclostationary (i.e., periodic) processes with arbitrary marginal distributions. Following the same naming convention with the initial publications, and since our approach uses as an auxiliary model the SMA scheme, the proposed method is termed Symmetric Moving Average (neaRly) To Anything (SMARTA). Alternatively, given that the latter schemes make use of the ICDF, which is generally a non-linear function, they can be viewed as a non-linear variation of underlying linear stochastic models (e.g., AR or SMA). The use of the ICDF in the abovementioned, Nataf-based, schemes ensures that the generated data will have the target distribution but on the other hand it is recognized that the Pearson correlation coefficient (which is used to express the dependencies in all linear stochastic models) is not invariant under such non-linear monotonic transformations (Embrechts et al., 1999). Therefore, the main challenge of such approaches, lies in identifying the "*equivalent*" correlations coefficients that should be used within the generation procedure (Gaussian domain) in order to attain the target correlation structure in the actual (i.e., real) domain. The latter relationship (i.e., that of equivalent and target correlations) can be expressed theoretically through a double infinite integral, which can be approximated with the use of numerical techniques such as the one employed herein.

Further details about the proposed approach can be found in sections 2 and 3, where the latter is further divided in four subsections. Particularly, section 2 presents some key concepts regarding modeling of auto-dependence structure in general; while subsections 3.1 and 3.2 develop the theoretical background of the proposed approach; next, subsection 3.3 describes the auxiliary SMA model and lastly, subsection 3.4 summarizes the overall approach and provides the generation mechanism of SMARTA in step-by-step manner. The generality of SMARTA is illustrated through a series of numerical examples, hypothetical (section 4) and real-world (section 5), including the simulation of both univariate and multivariate time series. Finally, in section 6 we synopsize and discuss the proposed modelling approach.

2 Modelling the auto-dependence structure

Before describing SMARTA, it is considered useful to provide a brief introduction to the tools that allow the mathematical description of the auto-dependence structure of a stochastic process. For a more thorough treatment, the interested reader is referred to the works of Papoulis (1991) and Lindgren et al. (2013). To elaborate, let $\underline{x}_t, t \in \mathbb{Z}$ be a discrete-time stationary process, indexed using t, with finite variance $\sigma^2 := \mathrm{Var}[\underline{x}_t]$ and autocorrelation function $\rho_\tau := \mathrm{Corr}[\underline{x}_t, \underline{x}_{t+\tau}] = \rho_{|\tau|}$, where $\tau = 0, \pm1, \pm2, \dots$ denotes the time lag. The autocovariance function (ACVF) of the process can be obtained by, $c_\tau := \mathrm{Cov}[\underline{x}_t, \underline{x}_{t+\tau}] = \sigma^2 \rho_\tau$. It is reminded that a valid autocorrelation structure has to be positive definite (e.g., Lindgren, 2013; Papoulis, 1991), which can be readily checked by formulating, and testing for positive definiteness, the so-called $(n \times n)$ autocorrelation matrix $\boldsymbol{R}$, whose $i^{\text{th}}, j^{\text{th}}$ elements are being determined by, $\boldsymbol{R}_{[i,j]} = \rho_{|i-j|}$. Besides the ACF and ACVF, another particularly useful stochastic tool, is the climacogram (CG, Koutsoyiannis, 2010, 2016), which is typically depicted using a log-log plot, and expresses the variance of the aggregated $\left(\underline{X}_l^{(k)}\right)$ or time averaged $\left(\underline{x}_l^{(k)}\right)$ process at scale $k \in \mathbb{Z}^+$. We point out that the notation employed herein slightly differs from the typical one, since we restrict our attention to discrete-time processes. Assuming that $\underline{x}_t$ denotes a discrete-time stationary process at the basic time scale $k = 1$, the discrete-time aggregated process at scale $k > 1$ can be obtained by,

$$\underline{X}_l^{(k)} \coloneqq \sum_{t=(l-1)k+1}^{kl} \underline{x}_t \tag{1}$$

while the averaged discrete-time process is obtained by, $\underline{x}_l^{(k)} = \underline{X}_l^{(k)}/k$. Hence, the corresponding climacograms of the discrete-time aggregated and averaged process can be defined as $\Gamma^{(k)} \coloneqq \mathrm{Var}\left[\underline{X}_l^{(k)}\right]$ and $\gamma^{(k)} \coloneqq \mathrm{Var}\left[\underline{x}_l^{(k)}\right]$ respectively. Moreover, as shown by Beran (1994 p. 3), as well as by Koutsoyiannis (2010, 2016), the variance over scales (i.e., the CG) and the ACVF (and therefore ACF) are interrelated. Specifically, if the theoretical ACVF (or ACF), c_τ at the basic time scale ($k = 1$) is known, the corresponding theoretical discrete-time climacogram of the aggregated process can be calculated using the following equation,

$$\Gamma^{(k)} = c_0 k + 2\sum_{\tau=1}^{k-1}(k-\tau)\, c_\tau \tag{2}$$

while the averaged one can be obtained by, $\gamma^{(k)} = \Gamma^{(k)}/k^2$. The recursive application of the following equation facilitates the calculation of the climacogram $\Gamma^{(k)}$,

$$\Gamma^{(k)} = 2\Gamma^{(k-1)} - \Gamma^{(k-2)} + 2c_{k-1} \tag{3}$$

It is noted that, $\Gamma^{(1)} = \gamma^{(1)} = c_0 = \sigma^2$, while $\Gamma^{(0)} = 0$. The inverse relationship that calculates the ACVF of the aggregated discrete-time process $\left(\underline{X}_l^{(k)}\right)$, denoted $C_\tau^{(k)} \coloneqq \mathrm{Cov}\left[\underline{X}_l^{(k)}, \underline{X}_{l+\tau}^{(k)}\right]$, at time scale k given the theoretical climacogram is given by (Koutsoyiannis, 2017),

$$C_\tau^{(k)} = \frac{\Gamma^{(|\tau+1|k)} + \Gamma^{(|\tau-1|k)}}{2} - \Gamma^{(|\tau|k)}, \qquad \tau \geq 0 \tag{4}$$

Furthermore, the ACVF, $C_\tau^{(k)}$ at scale k is linked with the ACVF, c_τ, of the basic time scale $k = 1$, through the following relationship,

$$C_\tau^{(k)} = \sum_{t=1}^{k}\sum_{r=\tau k+1}^{(1+\tau)k} \mathrm{Cov}[\underline{x}_t, \underline{x}_r] = \sum_{t=1}^{k}\sum_{r=\tau k+1}^{(1+\tau)k} c_{|t-r|}, \qquad \tau \geq 0 \tag{5}$$

Analogously, the ACVF of the time averaged discrete-time process $\left(\underline{x}_l^{(k)}\right)$ at scale k, denoted $c_\tau^{(k)} \coloneqq \mathrm{Cov}\left[\underline{x}_l^{(k)}, \underline{x}_{l+\tau}^{(k)}\right]$, is obtained by $c_\tau^{(k)} = C_\tau^{(k)}/k^2$. Hence, the ACF of the aggregated discrete-time process at time scale k can be obtained by $\rho_\tau^{(k)} = C_\tau^{(k)}/\Gamma^{(k)}$, while the ACF of the time averaged discrete-time process by $\rho_\tau^{(k)} = c_\tau^{(k)}/\gamma^{(k)}$. Note that the ACF of the aggregated and time averaged process are identical, due to standardization of the corresponding ACVF with the variance. It is also noted that $C_0^{(k)} = \Gamma^{(k)}$ and $C_\tau^{(1)} = c_\tau$, while similarly, $c_0^{(k)} = \gamma^{(k)}$ and $c_\tau^{(1)} = c_\tau$.

Undoubtedly, the most commonly-employed tool to characterize the auto-dependence structure is the autocorrelation function (ACF). The literature offers a plethora of theoretical models in both continuous and discrete time (Dimitriadis & Koutsoyiannis, 2015; Gneiting, 2000; Gneiting & Schlather, 2004; Koutsoyiannis, 2000, 2016; Papalexiou, 2018; Papalexiou et al., 2011), that can be easily combined with the proposed approach (see next section). In this work we use the discrete-time Cauchy-type autocorrelation structure (CAS) of Koutsoyiannis (2000) due to its simple and parsimonious form (a desired property in stochastic modelling), which however does not hinder its ability to model a wide range of short (ARMA-type) and long-range dependence structures (including HK behavior). CAS is a two-parameter power-type autocorrelation structure which, in its simplest form, if the ACF has constant and positive sign (as in the case of geophysical and hydrometeorological processes), is given by,

$$\rho_\tau^{\mathrm{CAS}} = (1 + \kappa\beta\tau)^{-1/\beta}, \qquad \tau \geq 0 \tag{6}$$

where $\beta \geq 0$ and $\kappa > 0$ are parameters that control the degree of dependence (or persistence) of the process. It is remarked that the autocorrelation function of an HK (i.e., fGn) process consists a special case (or a very good approximation) of the latter model (i.e., Eq. (6)) whose theoretical ACF is given by,

$$\rho_\tau^{\mathrm{HK}} = \frac{1}{2}\left(|\tau - 1|^{2H} - 2|\tau|^{2H} + |\tau + 1|^{2H}\right) \tag{7}$$

where H is the Hurst coefficient ($0 \leq H \leq 1$), which, loosely speaking, controls the degree of long-term dependence (or persistence) of the process. It has been shown that for large time lags and $H > 0.5$, the parameter β of CAS is related to the H coefficient of an HK ACF through the relationship $\beta = 1/(2 - 2H) > 1$, thus asymptotically resembling the right tail of latter theoretical model. More specifically, for $\beta > 1$ and when κ is set equal to κ_0, see Eq. (8), CAS closely approximates the theoretical ACF of an HK process, even for small time lags.

$$\kappa = \kappa_0 := \frac{1}{\beta\left[\left(1 - \frac{1}{\beta}\right)\left(1 - \frac{1}{2\beta}\right)\right]^{\beta}} \tag{8}$$

In addition, the ACF of an SRD process (ARMA-type) can be obtained through CAS, by setting $\beta = 0$ and applying the L' Hôpital's rule. The resulting SRD ACF is given by,

$$\rho_\tau^{\mathrm{SRD}} = \exp(-\kappa\tau) \tag{9}$$

Furthermore, when $\kappa = -\ln(\rho_1)$, and $0 \leq \rho_1 \leq 1$, Eq. (8) reduces to the classic Markovian ACF of an AR(1) process, given by, $\rho_\tau^{\mathrm{AR(1)}} = \rho_1^{|\tau|}$. For other parameter values, CAS resembles a plethora of alternative autocorrelation structures, that differ from the aforementioned classic models (for further details see, Koutsoyiannis, 2000). The flexibility of CAS is illustrated in **Figure 1**a where we depict (in a log-log scale) the theoretical ACF of various HK processes, characterized by different values of Hurst coefficient, H, as well as, their approximation with CAS. The close agreement of the two theoretical models is further validated in **Figure 1**b where we plot (also in log-log scale) their climacograms (assuming $\sigma^2 = c_0 = 1$), which are practically indistinguishable. It is noted that for an HK process, which exhibits simple and constant scaling laws, the slope s of the climacogram $\gamma^{(k)}$, i.e., the log-log derivative $s \coloneqq \mathrm{d}\big(\ln(\gamma^{(k)})\big)/\mathrm{d}(\ln(k))$, is related with H parameter by $s = 2H - 2$. The resemblance of the HK and CAS is confirmed by estimating the average mean square error (MSE) of the depicted processes by means of both ACF and climacogram. In terms of ACF, the average MSE value is 0.01 and the corresponding value in terms of climacogram is 0.66.

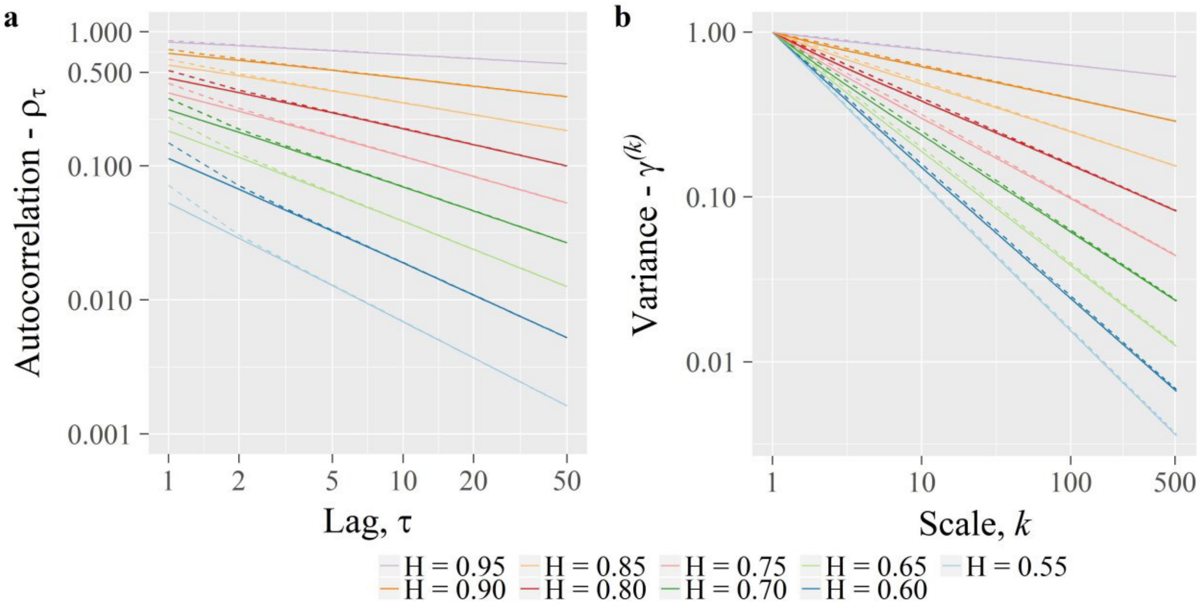

Figure 1. a) Autocorrelation functions and b) climacograms of HK processes exhibiting different Hurst coefficients (dashed lines) and their approximation with the CAS (continuous line).

Considering the practical aspects of the auto-dependence structure identification procedure (e.g., estimation of the parameters of CAS or any other theoretical structure, given a sample time series), it is remarked that it is a challenging task, due to the fact that the sample estimates of variance and autocorrelation coefficients (i.e., empirical variance and ACF - calculated from the historical time series) are negatively biased (e.g., Beran, 1994; Koutsoyiannis, 2003, 2016, 2017), especially in the presence of LRD (e.g., HK behavior). A thorough treatment on the subject lies beyond the scope of this study, as it has been extensively documented by the aforementioned authors, as well as by Dimitriadis and Koutsoyiannis (2015) who highlighted the advantages of using the climacogram, in comparison with the ACF and power spectrum, for the identification of the auto-dependence structure. The latter authors, via an extended analysis of a wide range of SRD and LRD processes, showed that the climacogram exhibits less uncertainty and bias in its estimation, which can be easily estimated *a priori*, thus providing an attractive alternative to the latter classic approaches. Further to this, the latter stochastic tool can be used as a basis for LRD identification algorithms (e.g., Tyralis & Koutsoyiannis, 2011), as well as for the development additional tools (e.g., the climacospectrum) that provide further insights regarding the asymptotic behavior of the process (Koutsoyiannis, 2016, 2017). It is noted that in this work, the above-mentioned stochastic tools (i.e., ACF and CG) are mainly employed for "*diagnostic*", and not for identification purposes, i.e., to verify that the simulated processes exhibit the desired dependence properties.

3 Methodology

3.1 Theoretical background of the SMARTA model

The central idea of the proposed approach is based on the Nataf's joint distribution model (NDM, Nataf, 1962) which has been originally implemented for the generation of cross-correlated, yet serially independent, random vectors with arbitrary distributions. One of its key assumptions, which consequently holds for SMARTA or any other Nataf-based method, is that the employed distributions owe to have finite variance. This assumption is implied throughout this work.

NDM gained popularity after the works of Liu and Der Kiureghian (1986) and Cario and Nelson (1997), who also coined the term NORmal To Anything (NORTA) procedure and also accounted for combinations of continuous and discrete marginal distributions. Its main concept lies in establishing joint relationships with the use of an auxiliary multivariate standard normal (i.e., Gaussian) distribution (using an appropriately adjusted correlation matrix); generating correlated standard normal variates and then mapping them to the actual domain using their ICDF. As noted by Cario and Nelson (1997) and further investigated by Lebrun and Dutfoy (2009), NDM is related to the Gaussian copula since the variables' joint distribution is established through the multivariate Gaussian distribution.

An interesting point concerning NDM (see, Tsoukalas et al., 2018a) is that it can be retrospectively associated with several well-known hydrological approaches (e.g., Kelly & Krzysztofowicz, 1997; Klemeš & Borůvka, 1974; Matalas, 1967). Among them, we distinguish the so-called Wilks' type weather generators (Wilks, 1998), which have motivated a significant amount of research during the last decades. The latter author, in an effort to simulate cross-correlated random variates, representing either the precipitation occurrence or amount process (neglecting temporal dependence), proposed the simulation of cross-correlated Gaussian variables and their subsequent mapping via their ICDF. Wilks empirically observed that a monotonic relationship exists which links the correlation coefficients of the Gaussian and "real" domain. Hence, the use of inflated correlation coefficients was proposed within the multivariate Gaussian distribution, in order to attain random variates with the required cross-correlation and distribution. The latter class of models is reviewed in the works of Wilks and Wilby (1999) and Ailliot et al., (2015).

In this study, we employ the concept of NDM, but in a different context, i.e., for the simulation of stationary any-range-dependent stochastic processes. Particularly, the rationale of NDM is combined with an auxiliary Gaussian process in order to capture the stochastic structure (in terms of autocorrelation and cross-correlation coefficients) of the target process and simultaneously preserve the desired marginal distributions after the use of the ICDF.

Suppose that the goal is to generate a m-dimensional discrete-time stationary process $\underline{\boldsymbol{x}}_t = \left[\underline{x}_t^1, \dots, \underline{x}_t^i, \dots, \underline{x}_t^m\right]^{\mathrm{T}}$, where t is the time index and the indices $i, j = 1, \dots, m$ are used to refer to individual process $\underline{x}_t^i$ and $\underline{x}_t^j$ respectively. Also let, $\boldsymbol{x}_t = \left[x_t^1, \dots, x_t^i, \dots, x_t^m\right]^{\mathrm{T}}$ denote a realization of it. Furthermore, let us assign a target cumulative distribution function (CDF), denoted by, $F_{\underline{x}^i} \coloneqq P\left(\underline{x}^i \le x\right)$ to each individual process $\underline{x}_t^i$, and let $\rho_{t,t+\tau}^{i,j} \coloneqq \mathrm{Corr}[\underline{x}_t^i, \underline{x}_{t+\tau}^j]$ denote the target Pearson's correlation coefficient between $\underline{x}_t^i$ and $\underline{x}_t^j$ for time lag τ.

Likewise, and using the same notation as above, let $\underline{\boldsymbol{z}}_t = \left[\underline{z}_t^1, \dots, \underline{z}_t^i, \dots, \underline{z}_t^m\right]^{\mathrm{T}}$ be an auxiliary m-dimensional stationary standard Gaussian process with zero mean and unit variance. Also, let $\tilde{\rho}_{t,t+\tau}^{i,j} \coloneqq \mathrm{Corr}[\underline{z}_t^i, \underline{z}_{t+\tau}^j]$ denote the Pearson's correlation coefficient of the auxiliary process between $\underline{z}_t^i$ and $\underline{z}_t^j$ for time lag τ, hereafter, referred to as equivalent correlation coefficient. It is noted that throughout the paper the superscripts or subscripts of $\rho_{t,t+\tau}^{i,j}$ or $\tilde{\rho}_{t,t+\tau}^{i,j}$ may be omitted when possible. For brevity, the target autocorrelation of the process $\underline{x}_t^i$ will be denoted ρ_τ^i and its lag-τ cross-correlation with $\underline{x}_t^j$ as $\rho_\tau^{i,j}$.

As mentioned earlier, the idea behind SMARTA lies in simulating an auxiliary standard Gaussian process $\underline{\boldsymbol{z}}_t$ using the SMA model with such parameters that after applying the inverse of their distribution function, results in a process $\underline{\boldsymbol{x}}_t$ with the desired correlation structure and marginal distributions. The latter operation can be written as follows,

$$\underline{x}_t^i = F_{\underline{x}^i}^{-1}\left(\Phi\left(\underline{z}_t^i\right)\right) \tag{10}$$

where $\Phi(\cdot)$ denotes the standard normal CDF and $F_{\underline{x}^i}^{-1}(\cdot)$ stands for the ICDF of process $\underline{x}_t^i$. An advantage of the above scheme is that since the ICDFs of the target distributions are employed (given that they can be analytically or numerically evaluated), the process $\underline{x}_t^i$ will inevitably have the desired marginal properties. On the other hand, the Pearson's correlation coefficient is not invariant under such non-linear monotonic transformations, hence $\rho_{t,t+\tau}^{i,j}$ will differ from $\tilde{\rho}_{t,t+\tau}^{i,j}$. However, as discussed in the literature, they are related (e.g., Biller & Nelson, 2003; Cario & Nelson, 1997; Der Kiureghian & Liu, 1986). Since Eq. (10) holds, we can write,

$$\rho_{t,t+\tau}^{i,j} = \mathrm{Corr}\left[\underline{x}_t^i, \underline{x}_{t+\tau}^j\right] = \mathrm{Corr}\left[F_{\underline{x}^i}^{-1}\left(\Phi\left(\underline{z}_t^i\right)\right), F_{\underline{x}^j}^{-1}\left(\Phi\left(\underline{z}_{t+\tau}^j\right)\right)\right] \tag{11}$$

Using the definition of Pearson's correlation coefficient, we can also write (for the sake of simplicity the time index t is omitted when possible due to stationarity),

$$\rho_{t,t+\tau}^{i,j} = \mathrm{Corr}\left[\underline{x}_t^i, \underline{x}_{t+\tau}^j\right] = \frac{\mathrm{E}\left[\underline{x}_t^i\, \underline{x}_{t+\tau}^j\right] - \mathrm{E}\left[\underline{x}^i\right] \mathrm{E}\left[\underline{x}^j\right]}{\sqrt{\mathrm{Var}\left[\underline{x}^i\right] \mathrm{Var}\left[\underline{x}^j\right]}} \tag{12}$$

where $\mathrm{E}\left[\underline{x}^i\right], \mathrm{E}\left[\underline{x}^j\right]$ and $\mathrm{Var}\left[\underline{x}^i\right], \mathrm{Var}\left[\underline{x}^j\right]$ denote the mean and variance of $\underline{x}^i$ and $\underline{x}^j$ respectively; which are known from the corresponding distributions $F_{\underline{x}^i}$ and $F_{\underline{x}^j}$ and have to be finite. Subsequently, using Eq. (10) and the first cross-product moment of $\underline{x}_t^i$ and $\underline{x}_{t+\tau}^j$ we obtain,

$$\begin{aligned} \mathrm{E}\left[\underline{x}_t^i \;\; \underline{x}_{t+\tau}^j\right] &= \mathrm{E}\left[F_{\underline{x}^i}^{-1}\left(\Phi\left(\underline{z}_t^i\right)\right) F_{\underline{x}^j}^{-1}\left(\Phi\left(\underline{z}_{t+\tau}^j\right)\right)\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{\underline{x}^i}^{-1}\left(\Phi\left(z_t^i\right)\right) F_{\underline{x}^j}^{-1}\left(\Phi\left(z_{t+\tau}^j\right)\right) \varphi_2\left(z_t^i, z_{t+\tau}^j, \tilde{\rho}_{t,t+\tau}^{i,j}\right) \mathrm{d}z_t^i \mathrm{d}z_{t+\tau}^j \end{aligned} \tag{13}$$

where $\varphi_2\left(z_t^i, z_{t+\tau}^j, \tilde{\rho}_{t,t+\tau}^{i,j}\right)$ is the bivariate standard normal probability density function. Hence, by substituting Eq. (13) to Eq. (11) we obtain,

$$\rho_{t,t+\tau}^{i,j} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{\underline{x}^i}^{-1}\left(\Phi\left(z_t^i\right)\right) F_{\underline{x}^j}^{-1}\left(\Phi\left(z_{t+\tau}^j\right)\right) \varphi_2\left(z_t^i, z_{t-\tau}^j, \tilde{\rho}_{t+\tau}^{i,j}\right) \mathrm{d}z_t^i \mathrm{d}z_{t+\tau}^j - \mathrm{E}\left[\underline{x}^i\right] \mathrm{E}\left[\underline{x}^j\right]}{\sqrt{\mathrm{Var}\left[\underline{x}^i\right] \mathrm{Var}\left[\underline{x}^j\right]}} \tag{14}$$

Inspection of Eq. (14) indicates that $\rho_{t,t+\tau}^{i,j}$ is a function of $\tilde{\rho}_{t,t+\tau}^{i,j}$, since all other quantities are already known from the target (i.e., given) distributions $F_{\underline{x}^i}$ and $F_{\underline{x}^j}$. Therefore, it is compactly written as,

$$\rho_{t,t+\tau}^{i,j} = \mathcal{F}\left(\tilde{\rho}_{t,t+\tau}^{i,j} \middle| F_{\underline{x}^i}, F_{\underline{x}^j}\right) \tag{15}$$

where $\mathcal{F}(\cdot)$ is an abbreviation of the function defined by Eq. (14).

This relationship implies that prior to the estimation of the auxiliary model's parameters it is essential to identify, and next use within parameter estimation, the equivalent correlations, $\tilde{\rho}_{t,t+\tau}^{i,j}$, that result to the target correlations, $\rho_{t,t+\tau}^{i,j}$, after the subsequent mapping of the auxiliary process to the actual domain. This can be achieved through inversion of Eq. (15), i.e., $\tilde{\rho}_{t,t+\tau}^{i,j} = \mathcal{F}^{-1}\left(\rho_{t,t+\tau}^{i,j} \middle| F_{\underline{x}^i}, F_{\underline{x}^j}\right)$.

3.2 Identification of equivalent correlation coefficients

Provided that the identification of equivalent correlation coefficients can be accomplished on a pairwise basis, and for the sake of simplicity, let us define $\underline{x}_\xi \coloneqq \underline{x}_t^i$ and $\underline{x}_\psi \coloneqq \underline{x}_{t+\tau}^j$, hence

$\tilde{\rho}_{\xi,\psi}$ and $\rho_{\xi,\psi}$ stand for the equivalent and the target correlation coefficients respectively. Furthermore, let $F_{\underline{x}_\xi}$ and $F_{\underline{x}_\psi}$ denote the corresponding target distributions. It is reminded that our ultimate objective is to establish a relationship between $\tilde{\rho}_{\xi,\psi}$ and $\rho_{\xi,\psi}$ and eventually find the appropriate value of $\tilde{\rho}_{\xi,\psi}$ that results in the target correlation $\rho_{\xi,\psi}$ after the mapping operation of Eq. (10). It is acknowledged that Eq. (15) does not have a general closed-form solution, with the exception of few special cases, hence it is typically identified via numerical techniques such as crude search, quadrature methods as well as Monte-Carlo procedures (Cario & Nelson, 1996, 1997; Chen, 2001; Li & Hammond, 1975; Liu & Der Kiureghian, 1986; Xiao, 2014). The abovementioned authors provided a series of Lemmas that can be used in order to establish the relationship of Eq. (15). Among them,

Lemma 1. $\rho_{\xi,\psi}$ is a strictly increasing function of $\tilde{\rho}_{\xi,\psi}$.

Lemma 2. $\tilde{\rho}_{\xi,\psi} = 0$ for $\rho_{\xi,\psi} = 0$ as well as, $\tilde{\rho}_{\xi,\psi} \geq (\leq)\ 0$ if $\rho_{\xi,\psi} \geq (\leq)\ 0$.

Lemma 3. $|\rho_{\xi,\psi}| \leq |\tilde{\rho}_{\xi,\psi}|$.

It is remarked that the equality sign in Lemma 3 is valid when $\rho_{\xi,\psi} = 0$ or when both marginals are Gaussian. Furthermore, the minimum and maximum attainable values of $\rho_{\xi,\psi}$ are in accordance with the Fréchet-Hoeffding bounds (Fréchet, 1957; Hoeffding, 1994) and are given for $\tilde{\rho}_{\xi,\psi} = -1$ and $\tilde{\rho}_{\xi,\psi} = 1$, respectively. Particularly the following relationship holds true, $-1 \leq \mathcal{F}\left(-1 \middle| F_{\underline{x}_\xi}, F_{\underline{x}_\psi}\right) \leq \rho_{\xi,\psi} \leq \mathcal{F}\left(1 \middle| F_{\underline{x}_\xi}, F_{\underline{x}_\psi}\right) \leq 1$. See also the work of Whitt (1976) for a comprehensive discussion on the topic. In this paper, unless stated otherwise, in order to establish the relationship of Eq. (15) we employ the simple, yet efficient method proposed by Tsoukalas et al., (2018a), which in a nutshell, is based on the evaluation of few pairs of $\rho_{\xi,\psi}$ and $\tilde{\rho}_{\xi,\psi}$ using Monte-Carlo simulation and subsequently, the establishment of the relationship of Eq. (15) through polynomial interpolation (see also, Appendix A).

3.2.1 An illustrative example

To shed some light on the functional form of $\mathcal{F}(\cdot)$ let us consider the case where both variables $\underline{x}_\xi$ and $\underline{x}_\psi$ are described by the two-parameter Gamma distribution ($\mathcal{G}$). The probability density function (PDF) of the latter distribution is given by,

$$f_{\mathcal{G}}(x; a, b) = \frac{1}{|b|\Gamma(a)}\left(\frac{x}{b}\right)^{a-1} \exp\left(-\frac{x}{b}\right), \qquad x > 0 \tag{16}$$

where $\Gamma(\cdot)$ denotes the gamma function and $a > 0$ and $b \neq 0$ are shape and scale parameters, respectively. **Figure 2**a depicts the relationship among $\tilde{\rho}_{\xi,\psi}$ and $\rho_{\xi,\psi}$ (i.e., $\mathcal{F}(\cdot)$; computed via numerical integration) for various values of distribution parameters. Specifically, we assumed $a \coloneqq a_\xi = a_\psi$ and constant $b \coloneqq b_\xi = b_\psi = 1$. We remind that the theoretical skewness coefficient of a Gamma distributed variable is given by $C_{s_{\underline{x}}} = 2/\sqrt{a}$. From the latter figure we observe that the non-linearity of $\mathcal{F}(\cdot)$ increases with low values of a (i.e., high skewness), and that the maximum attainable value of $\rho_{\xi,\psi}$ is equal to 1, due to the fact that $F_{\underline{x}_\xi} \equiv F_{\underline{x}_\psi}$. In addition, one may observe that the shape parameter a is also related to the minimum attainable value of $\rho_{\xi,\psi}$. For example, when $a = 0.01$ the latter value is practically restricted to zero, something that may be considered a reasonable behavior, attributed to the very high value of positive skewness which does not allow for negative correlations. In a similar vein, in **Figure 2**b we set $a_\xi = 5$ and vary parameter a_ψ from 5 to 0.01 (assuming again that $b \coloneqq b_\xi = b_\psi = 1$). In this case, both the minimum and maximum attainable values of $\rho_{\xi,\psi}$ are affected. It is observed that, when a_ξ and a_ψ exhibit significant differences, the range of feasible values $\rho_{\xi,\psi}$ is getting narrower. This implies that two variables with considerable different shape (expressed through parameter a) cannot be highly correlated. From an engineering point of view, and similar to the previous case (i.e., when $a \coloneqq a_\xi = a_\psi$), this is barely considered a limitation of the proposed approach, since such behavior is rarely encountered in hydrometeorological processes. For instance, it is not expected, or rational, two processes, one with skewness ~0.9 and one with 20 to be highly correlated (positively or negatively). In any case, we stress the importance of checking the range of attainable correlation coefficients when employing the concept of NDM, (see, Demirtas & Hedeker, 2011; Leonov & Qaqish, 2017), especially within the context of stochastic process simulation. For instance, given the non-linear and asymmetric nature of $\mathcal{F}(\cdot)$, for some combinations of marginal distributions, a target correlation coefficient may be inadmissible. This constraint, and the fact that the target marginal distributions ought to have finite variance, drove us to add the designation "nearly" when naming the method. However, in the examples employed in this work, such problems did not occur (for a simulation example also involving negative cross-correlations see section 4.2), a fact which by no means overrules the aforementioned need for compatibility verification.

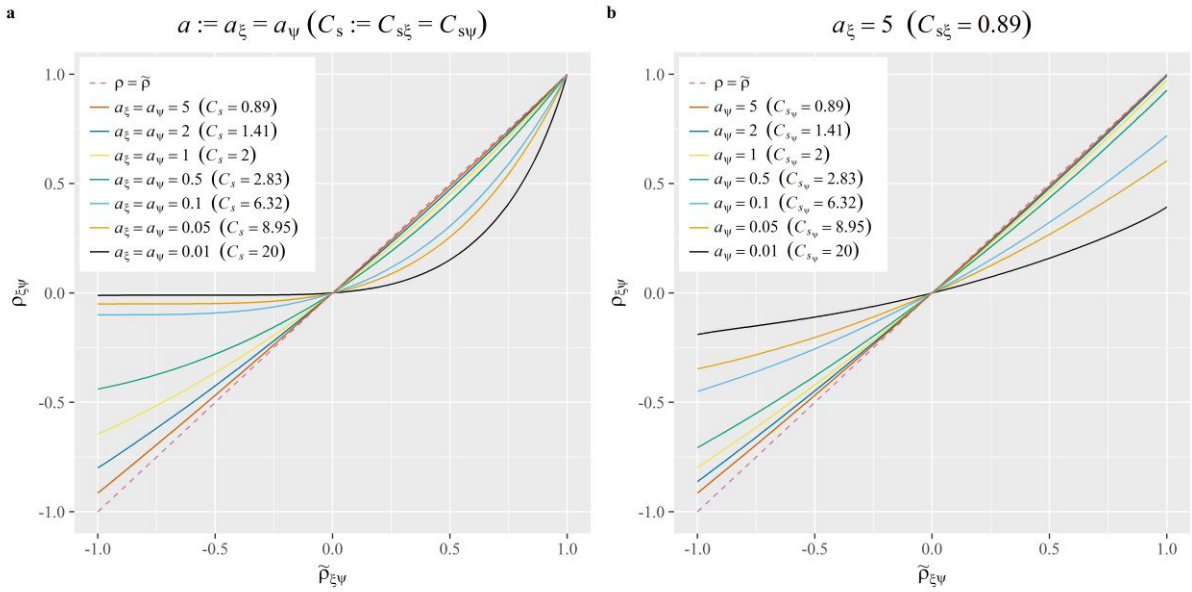

Figure 2. Graphical illustration of function $\mathcal{F}(\cdot)$ (see, Eq. (15)) that expresses the relationship between the equivalent, $\tilde{\rho}_{\xi,\psi}$ and target $\rho_{\xi,\psi}$ correlation coefficients assuming that both $\underline{x}_\xi$ and $\underline{x}_\psi$ are described by the two-parameter Gamma distribution (assuming that $b := b_\xi = b_\psi = 1$) with a) equal shape parameters ($\text{i.e.}, a := a_\xi = a_\psi$) and b) different shape parameters by setting $a_\xi = 5$ and varying a_ψ from 5 to 0.01.

Evidently, the proper and accurate identification of the relationship $\mathcal{F}(\cdot)$ has a crucial role in NDM-based schemes, since its misspecification may lead to simulation errors. Hence, to assess the suitability of the algorithm of Appendix A, which is extensively used in this work, we employed the latter and recreated the cases depicted in **Figure 2**; which concerned the identification of equivalent correlation coefficients of two Gamma-distributed variables for various values of shape parameters. After the specification of the relationship $\mathcal{F}(\cdot)$ by the latter algorithm, the target correlations where evaluated for values of $\tilde{\rho}_{\xi,\psi} \in [-1,1]$ sampled by 0.01. To provide a quantitative comparison, we estimated the MSE and maximum square error (Max(SE)) between the estimates of the numerical integration method (i.e., **Figure 2**) and those of the aforementioned algorithm. A synopsis of the results is given on **Table 1**, where the panels (a) and (b) corresponds to those of **Figure 2**. The latter analysis illustrates the potential of the employed method to resemble the asymmetric and non-linear nature of $\mathcal{F}(\cdot)$ with high accuracy.

Table 1. Comparison between numerical integration and the algorithm of Appendix A for the numerical example illustrate in **Figure 2**. Panels a) and b) correspond to those of **Figure 2**.

a)

$a \coloneqq a_\xi = a_\psi \mid b := b_\xi = b_\psi = 1$		
Shape (a)	MSE	Max(SE)
0.01	8.03×10^{-5}	7.75×10^{-4}
0.05	5.81×10^{-5}	3.08×10^{-4}
0.1	2.44×10^{-6}	9.89×10^{-6}
0.5	4.33×10^{-6}	1.59×10^{-5}
1	3.31×10^{-6}	1.88×10^{-5}
2	1.22×10^{-6}	8.47×10^{-6}
5	3.70×10^{-6}	1.80×10^{-5}

b)

$a_\xi = 5 \mid b := b_\xi = b_\psi = 1$		
Shape (a_ψ)	MSE	Max(SE)
0.01	2.12×10^{-5}	3.79×10^{-4}
0.05	6.46×10^{-6}	2.70×10^{-5}
0.1	6.26×10^{-6}	4.15×10^{-5}
0.5	1.51×10^{-5}	9.37×10^{-5}
1	2.54×10^{-6}	1.13×10^{-5}
2	7.19×10^{-7}	3.20×10^{-6}
5	5.24×10^{-7}	1.77×10^{-6}

3.2.2 The Log-Normal case

As mentioned earlier, there are some exceptions that have a closed-form solution. Among them the Log-Normal case, which is of particular interest from a hydrological perspective. The PDF of the 3-parameter Log-Normal distribution ($\mathcal{LN}$) is given by,

$$f_{\mathcal{LN}}(x; a, b, c) = \frac{1}{(x-c)a\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(x-c)-b}{a}\right)^2\right), \qquad x > c \tag{17}$$

where $a > 0$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$ denote the shape, scale and location parameters respectively; while, when $c = 0$, the distribution reduces to the 2-parameter Log-Normal distribution. As shown in Mostafa and Mahmoud (1964), yet without direct reference to NDM, for two random variables $\underline{x}_\xi$ and $\underline{x}_\psi$ that are Log-Normally distributed, Eq. (14) simplifies to,

$$\rho_{\xi,\psi} = \frac{\exp\left(\tilde{\rho}_{\xi,\psi} a_\xi a_\psi\right) - 1}{\sqrt{\left(\exp(a_\xi^2) - 1\right)\left(\exp(a_\psi^2) - 1\right)}} \tag{18}$$

Which can be easily inverted in order to directly provide the equivalent correlation coefficient $\tilde{\rho}_{\xi,\psi}$, given the target value of $\rho_{\xi,\psi}$. i.e.,

$$\tilde{\rho}_{\xi,\psi} = \frac{\mathrm{Ln}\left(1 + \rho_{\xi,\psi}\sqrt{\left(\exp(a_\xi^2) - 1\right)\left(\exp(a_\psi^2) - 1\right)}\right)}{a_\xi a_\psi} \tag{19}$$

It is worth remarking that Eq. (18) is identical with the one employed in the celebrated multivariate lag-1 Log-Normal model of Matalas (1967), in order to adjust the correlation coefficients, which interestingly can be identified as a Nataf-based approach.

3.2.3 A cautionary note

A delicate point worth standing concerns the use of alternative, rank-based dependence measures, such as Spearman's r_s and Kendall's t, for the parameter identification of NDM (or Gaussian copula). Under the assumption that both marginal distributions and copula are Gaussian (or more generally elliptical distributions), there is a one-to-one relationship between the aforementioned dependence measures and Pearson's correlation coefficient (ρ), which can be expressed as (e.g., Embrechts et al., 1999; Esscher, 1924; Kruskal, 1958; Lebrun & Dutfoy, 2009) (notice that the indices have been omitted for the sake of simplicity),

$$\rho = 2\sin\left(\frac{\pi r_s}{6}\right) \leftrightarrow r_s = \left(\frac{6}{\pi}\right)\arcsin\left(\frac{\rho}{2}\right) \tag{20}$$

$$\rho = \sin\left(\frac{\pi t}{2}\right) \leftrightarrow t = \left(\frac{2}{\pi}\right)\arcsin(\rho) \tag{21}$$

Both r_s and t are measures of concordance and are invariant to non-linear monotonic transformations (such as those imposed by Eq. (10)). Thus, specifying NDM with estimates of Pearson's correlation based on the conversion of empirical estimates of r_s or t will inevitably preserve the target values of r_s or t after the application of the mapping procedure (due to the property of invariance) but it will lead to misspecification of the underlying model (i.e., NDM) due to Eq. (14), and of course the target values of ρ won't be preserved.

3.3 The auxiliary SMA model

Having described the theoretical background of the proposed approach, this section provides a brief introduction to the univariate and multivariate Symmetric Moving Average (SMA) model of Koutsoyiannis (2000), which is used within SMARTA as an auxiliary standard Gaussian process. SMA model consists as a special case of the Backward-Forward Moving Average (BFMA) model, whose key idea is that a stochastic process $\underline{z}_t$ can be described as a weighted sum of infinite backward and forward random variables. Note that the notation slightly differs from the original one, in order to highlight the fact that the model is employed in the Gaussian domain using the equivalent correlation coefficients $\tilde{\rho}$, instead of the target correlation coefficients, ρ.

3.3.1 Univariate model

In practice, the SMA model slightly relaxes the assumptions of BFMA model and assumes that a stochastic process $\underline{z}_t$ can be described as a weighted sum of a finite number of backward and forward random variables. Particularly, the generating mechanism of the SMA model is given by the following equation,

$$\underline{z}_t = \sum_{\zeta=-q}^{q} \tilde{a}_{|\zeta|} \underline{v}_{t+\zeta} = \tilde{a}_q \underline{v}_{t-q} + \cdots + \tilde{a}_1 \underline{v}_{t-1} + \tilde{a}_0 \underline{v}_t + \tilde{a}_1 \underline{v}_{t+1} + \cdots + \tilde{a}_q \underline{v}_{t+q} \qquad (22)$$

where $\underline{v}_t$ are standard normal i.i.d. variables and $\tilde{a}_\zeta$ are internal model parameters (i.e., weight coefficients) that are assumed to be symmetric, i.e., $\tilde{a}_\zeta = \tilde{a}_{-\zeta}$ (for $\zeta = 1, 2, \ldots$) and approach zero after some value $|\zeta| > q$, where q denotes a large positive integer value. The selection of q depends on the degree of auto-dependence imposed by the target process (see Eq. (23)) and the desired level of accuracy. Furthermore, q cannot be greater than the length of the time series to simulate. Particularly, the parameters $\tilde{a}_\zeta$ are related to the autocorrelation coefficients $\tilde{\rho}_\tau$ via a $2q + 1$ equation system of the following form,

$$\tilde{\rho}_\tau = \sum_{\zeta=-q}^{q-\tau} \tilde{a}_{|\zeta|} \tilde{a}_{|\tau+\zeta|}, \qquad \tau = 0, 1, 2, \ldots, q \qquad (23)$$

$$\tilde{\rho}_\tau = \sum_{\zeta=\tau-q}^{q} \tilde{a}_\zeta \tilde{a}_{\tau-\zeta}, \qquad \tau = q + 1, \ldots, 2q \qquad (24)$$

Evidently, if Eq. (23) is honored, the model resembles the theoretical ACF up to $\tilde{\rho}_q$, while it decays to zero after $2q$ (see Eq. (24)). In order to estimate the parameters $\tilde{a}_\zeta$, Koutsoyiannis (2000) proposed two solutions, one closed-form and one based on a formulation of an optimization problem. The interested reader is referred to the latter publication for a thorough and in-depth description of the two methods. In this work we restrict our attention in briefly describing only the first one, since it is a fast and direct method. The aforementioned author showed that the discrete Fourier transformation (DFT) of $\tilde{a}_\zeta$, i.e., $S_{\tilde{a}}(\omega)$, is related to the power spectrum of the autocorrelation function, i.e., $S_{\tilde{\rho}}(\omega)$, by, $S_{\tilde{a}}(\omega) = \sqrt{2S_{\tilde{\rho}}(\omega)}$.
If the autocorrelation structure $\tilde{\rho}_\tau$ is known (or specified), its power spectrum can be calculated using the DFT, hence estimate $S_{\tilde{a}}(\omega)$. Then, by applying the inverse Fourier transformation

one can obtain the parameters $\tilde{a}_\zeta$. It is remarked that algorithms that facilitate the latter calculations are nowadays built-in in many high-level programming languages (e.g., R or MATLAB), which in turn allow the straightforward implementation of SMA and SMARTA models in most computational environments. At this point we note that Koutsoyiannis (2002, 2016) proposed an even simpler and straightforward procedure for the estimation of $\tilde{a}_\zeta$ coefficients, which however is applicable only for HK (i.e., fGn) type autocorrelation structures.

3.3.2 Multivariate model

Furthermore, the SMA model can be extended for the multivariate simulation of contemporaneously cross-correlated processes, via the explicit preservation of the lag-0 cross-correlation coefficients. This assumption, which significantly simplifies the parameter estimation procedure, is often regarded adequate within hydrological domain, and can be found in several other stochastic simulation schemes (e.g., Camacho et al., 1985; Efstratiadis et al., 2014; Koutsoyiannis & Manetas, 1996; Pegram & James, 1972; Tsoukalas et al., 2018a). With this in mind, for simulation of hydrometeorological processes characterized by strongly lagged cross-correlations (e.g., rainfall-runoff at fine time scales), it may be advantageous to employ the same modelling strategy as the one proposed herein, using alternative auxiliary Gaussian models that, apart from the lag-0 cross-correlations, are able to directly model (preferably, for parsimony and stability, in combination with suitable theoretical auto- and cross-correlation structures; e.g., similar to CAS) the lagged cross-correlation coefficients.

Regarding the multivariate SMA model, let $\underline{\boldsymbol{z}}_t = \left[\underline{z}_t^1, \ldots, \underline{z}_t^i, \ldots, \underline{z}_t^m\right]^{\mathrm{T}}$ be a m-dimensional vector, as defined in section 2, and $\tilde{\rho}_\tau^{i,j} \coloneqq \mathrm{Corr}\left[\underline{z}_t^i, \underline{z}_{t+\tau}^j\right]$ denote the equivalent lag-τ cross-correlation between processes $\underline{z}_t^i$ and $\underline{z}_t^j$ for time lag τ. Similar to the univariate case, each process $\underline{z}_t^i$ is represented by a weighted sum of random variables $\underline{v}_t^i$, i.e.,

$$\underline{z}_t^i = \sum_{\zeta=-q}^{q} \tilde{a}_{|\zeta|}^i \underline{v}_{t+\zeta}^i \tag{25}$$

In this case, the random variables $\underline{v}_t^i$ are considered serially independent but contemporaneously cross-correlated. Therefore, the problem lies in generating such variables in a way that they reproduce the equivalent lag-0 cross-correlation coefficients ($\tilde{\rho}_0^{i,j}$). It has been shown that it suffices to generate random variables $\underline{v}_t^i$ with correlation $\tilde{g}^{i,j} \coloneqq \mathrm{Corr}\left[\underline{v}_t^i, \underline{v}_t^j\right]$ equal to,

$$\tilde{g}^{i,j} = \frac{\tilde{\rho}_0^{i,j}}{\sum_{\zeta=-q}^{q} \tilde{a}_{|\zeta|}^i \tilde{a}_{|\zeta|}^j} \tag{26}$$

Hence, the ($m \times m$) correlation matrix $\widetilde{\boldsymbol{G}}$ is formulated, with ones in the diagonal and its $i^{\text{th}} \neq j^{\text{th}}$ elements determined by, $\widetilde{\boldsymbol{G}}_{[i,j]} = \tilde{g}^{i,j}$. Furthermore, the theoretical lag-τ cross-correlation structure (for $\tau = 0,1,2, \ldots$) of the model is given by,

$$\tilde{\rho}_\tau^{i,j} = \tilde{\rho}_0^{i,j} \frac{\sum_{\zeta=-q}^{q-\tau} \tilde{a}_{|\tau+\zeta|}^i \tilde{a}_{|\zeta|}^j}{\sum_{\zeta=-q}^{q} \tilde{a}_{|\zeta|}^i \tilde{a}_{|\zeta|}^j} = \tilde{g}^{i,j} \sum_{\zeta=-q}^{q-\tau} \tilde{a}_{|\tau+\zeta|}^i \tilde{a}_{|\zeta|}^j \tag{27}$$

Regarding simulation, a vector of correlated random variables $\underline{\boldsymbol{v}}_t = \left[\underline{v}_t^1, \ldots, \underline{v}_t^i, \ldots, \underline{v}_t^m\right]^{\mathrm{T}}$ can be generated by, $\underline{\boldsymbol{v}}_t = \widetilde{\boldsymbol{B}} \underline{\boldsymbol{w}}_t$, where $\underline{\boldsymbol{w}}_t = \left[\underline{w}_t^1, \ldots, \underline{w}_t^i \ldots, \underline{w}_t^m\right]^{\mathrm{T}}$ is a vector of standard normal i.i.d. variables, and $\widetilde{\boldsymbol{B}}$ is a $m \times m$ matrix obtained by finding the so-called square root of matrix $\widetilde{\boldsymbol{G}}$, i.e., Eq. (28). A solution to the latter problem can be obtained by standard decomposition techniques (e.g., Cholesky or singular value decomposition) or via optimization-based methods (Higham, 2002; Koutsoyiannis, 1999).

$$\tilde{\boldsymbol{B}}\tilde{\boldsymbol{B}}^{\mathrm{T}} = \tilde{\boldsymbol{G}} \tag{28}$$

In more detail, it is reminded that if $\tilde{\boldsymbol{G}}$ is positive definite (which indicates that the multivariate process is admissible), then Eq. (28) has infinite solutions, hence, both decomposition and optimization-based methods can be employed. On the other hand, when $\tilde{\boldsymbol{G}}$ is non-positive definite (implying that the multivariate process is inadmissible), the decomposition methods cannot offer a solution. In this case, optimization-based techniques can provide a potential remedy, by formulating an optimization problem, where the objective is to identify a matrix $\tilde{\boldsymbol{B}}^{*}$ which results to a feasible and near-to-optimum matrix $\tilde{\boldsymbol{G}}^{*} \coloneqq \tilde{\boldsymbol{B}}^{*}\, \tilde{\boldsymbol{B}}^{*\mathrm{T}}$ which is as closest (typically quantified in terms of some distance measure; e.g., Euclidean norm) as possible to the original matrix $\tilde{\boldsymbol{G}}$. Of course, in such cases, the target process will not be exactly resembled, while, the difference between $\tilde{\boldsymbol{G}}$ and $\tilde{\boldsymbol{G}}^{*}$ can be regarded as a proxy for the magnitude of approximation introduced to the simulation. Bras and Rodríguez-Iturbe (1985 p. 98), as well as Koutsoyiannis (1999) discuss several situations which may lead to a non-positive definite matrix $\tilde{\boldsymbol{G}}$. Almost all of these situations are related with the estimates of correlation coefficients from the empirical data. In the case of SMARTA, and provided that a feasible autocorrelation structure has been identified for each individual process, a non-positive definite matrix $\tilde{\boldsymbol{G}}$ may arise due to data-based estimates of lag-0 cross-correlation coefficients, imprecise approximation of equivalent correlation coefficients or incompatible combinations of marginal distributions, autocorrelation structures and target cross-correlations (see section 3.2.1). For instance, since the proposed scheme (in multivariate mode) treats each individual process separately of the cross-correlations, the simulation of highly cross-correlated processes with particularly different distributions and autocorrelation structures (e.g., very fast-decaying and very slow-decaying) may be infeasible (see section 4.2 for a simulation example involving both positively and negative cross-correlated LRD and SRD processes), even if the latter are individually valid.

At this point it is noted that an incidental contribution of SMARTA is the alleviation of a burden related to preservation of the skewness coefficient. As mentioned in the introduction, a broad class of linear stochastic models, in an attempt to preserve the coefficients of skewness of the target process, $\underline{x}_t$, employ non-Gaussian white noise for the innovation term, $\underline{v}_t$, typically from Pearson type-III distribution. However, the latter practice may lead to very high coefficients of skewness for the innovation term which are hardly attainable (Koutsoyiannis, 1999; Todini, 1980). This practice was also adopted by Koutsoyiannis (2000) in the original SMA scheme, where the Pearson type-III distribution has been employed for the generation of skewed white noise. More specifically, regarding the univariate formulation of the latter model (assuming $q = 2^{10}$), in **Figure 3**a-b we depict (from two distinct points of view) the relationship between the skewness coefficient $\left(C_{s_{\underline{v}}}\right)$ of innovation term, $\underline{v}_t$, that is required to attain the target coefficient of skewness $\left(C_{s_{\underline{x}}}\right)$ of the variable, $\underline{x}_t$, for several hypothetical HK process characterized by different values of H coefficient. See also Eq. (29) in Koutsoyiannis (2000). It is apparent from in **Figure 3**a-b that the higher the value of H, the higher the required skewness of the innovation term, $\underline{v}_t$. For example, in an HK process with $H = 0.8$, the skewness coefficient of innovation term $\underline{v}_t$ has to be set twice as high as than the one of $\underline{x}_t$. We remark that this issue is further amplified (not shown herein) when the underlying model is used in multivariate mode (Koutsoyiannis, 1999). On the other hand, SMARTA completely alleviates the latter difficulties since the SMA scheme is used as an auxiliary model in the standard

Normal (i.e., Gaussian) domain and the generated data are subsequently mapped to the actual domain using the target ICDFs. Therefore, the target marginal statistics are attained without making any attempts to generate skewed innovation terms, neither in univariate nor in multivariate mode. Moreover, an additional contribution of SMARTA regards the optimization problem that arises when the matrix $\tilde{\boldsymbol{G}}$ is non-positive. Particularly, the latter is simplified in a nearest correlation matrix problem, since the 3rd term of Eq. (28) in Koutsoyiannis (1999), that accounts for skewness, is no longer needed.

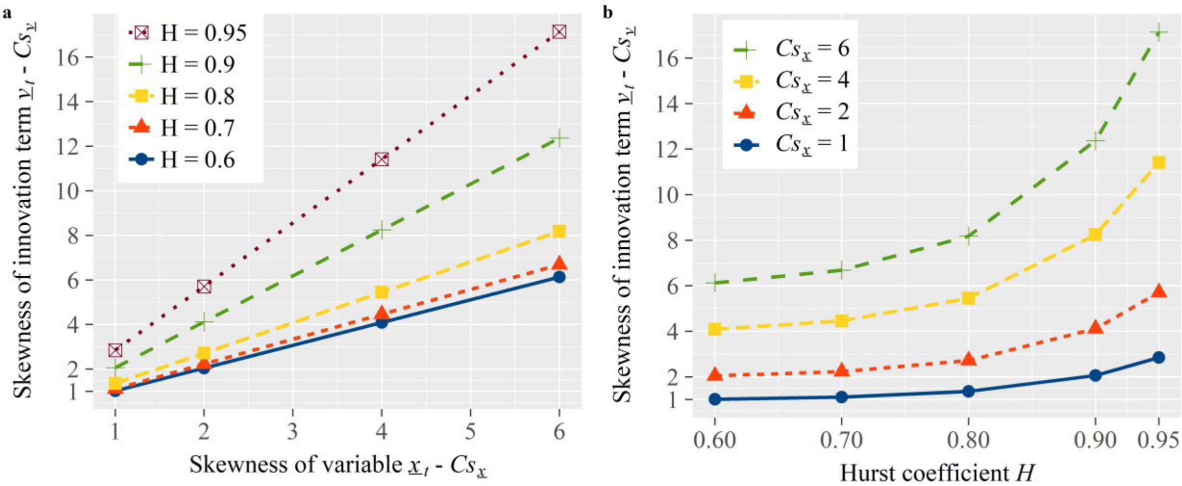

Figure 3. Graphical illustration of the relationship between the required skewness coefficient $\left(C_{s_{\underline{v}}}\right)$ of innovation term $\underline{v}_t$ and a) the skewness $\left(C_{s_{\underline{x}}}\right)$ of an fGn process $\underline{x}_t$ for various values of H and b) the value of H of an fGn process $\underline{x}_t$ for various values of skewness of $C_{s_{\underline{x}}}$ (using the SMA model with $q = 2^{10}$).

3.4 Generation procedure of SMARTA

Having described in detail all the key components of SMARTA approach in the previous sections, it is useful to provide the complete generation procedure, decomposed into the following six steps:

Step 1. Define a target distribution $F_{\underline{x}^i}$ for each process $\underline{x}_t^i$; $i = 1, \dots, m$. SMARTA, as well as all Nataf-based methods, is flexible in terms of distribution fitting method; hence one can select a fitting method of their preference.

Step 2. Define a target auto-correlation structure (ρ_τ^i) for each process $\underline{x}_t^i$; $i = 1, \dots, m$ using a theoretical ACF model. For instance, for each process $\underline{x}_t^i$ identify the parameters of CAS that better fit the observed data. Furthermore, in the multivariate case, identify the target lag-0 cross-correlation coefficients $(\rho_0^{i,j})$ between processes, $\underline{x}_t^i$ and $\underline{x}_t^j$; $i \neq j = 1, \dots, m$.

Step 3. Identify the equivalent correlation coefficients $(\tilde{\rho}_\tau^i)$ of each theoretical ACF, up to the maximum specified lag (which depends on the type of the process; LRD or SRD), for each process $\underline{x}_t^i$; $i = 1, \dots, m$. Furthermore, in the multivariate case, estimate the equivalent lag-0 cross-correlation coefficient $\tilde{\rho}_0^{i,j}$. Assuming that the algorithm of Appendix A is employed for the identification of equivalent correlations, and given the fact that it allows the direct estimation of the equivalent ACF up to any lag, the latter has to be employed m times, one for each process $\underline{x}_t^i$; $i = 1, \dots, m$. Furthermore, in order to estimate the lag-0 equivalent cross-correlation coefficient $\tilde{\rho}_0^{i,j}$, the same procedure should be employed $m(m-1)/2$ additional times. For instance, in a 4-dimensional problem ($m = 4$), the algorithm of Appendix A is executed in total, $m(m+1)/2$ times (=10).

Step 4. Calculate the parameters of the auxiliary SMA model (section 3.3), i.e., the weight coefficients ($\tilde{a}_{\zeta}^{i}$) of each auxiliary process $\underline{z}_t^i$; $i = 1, \dots, m$. Additionally, in the multivariate case, calculate the elements of matrices $\widetilde{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{B}}$ (see also, Eq. (26) and (28)).

Step 5. Employ the auxiliary Gaussian SMA model and generate a realization of the auxiliary univariate ($\underline{z}_t$) or multivariate process ($\underline{\boldsymbol{z}}_t$).

Step 6. Attain the actual process $\underline{x}_t$ (or $\underline{\boldsymbol{x}}_t$), by mapping the auxiliary Gaussian process $\underline{z}_t$ (or $\underline{\boldsymbol{z}}_t$) to the actual domain using the ICDF, $F_{\underline{x}^i}^{-1}$, of each process $\underline{x}_t^i$; $i = 1, \dots, m$, via Eq. (10).

By now, it should be clear that the basis of the proposed methodology consists an explicit simulation method, in terms of reproducing the distribution function (relieved from the limitations and constraints of such schemes; see section 1), that fundamentally differs from the other two typical schemes (implicit and transformation-based; see section 1) used in hydrology, which also employ linear stochastic models. Compared to the implicit approaches, that employ non-Gaussian white noise, Nataf-based schemes (e.g., SMARTA) alleviate several notable limitations. Among them, the approximation of the distribution function, the generation of negative values, the bounded dependence patterns and the (often) narrow type of possible correlation structures, which can be attributed to the limited number of schemes for which analytical equations can be derived to link the moments of the process with those of the white noise. Additionally, in contrast to transformation-based approaches, that aim to *normalize* the data, Nataf-based schemes explicitly model them using target marginal distributions. Though, it has to be noted, that in principle, the rationale of transformation-based approaches can be easily aligned with the theoretical background of Nataf's distribution model, by using the concept of equivalent (i.e., adjusted) correlation coefficients. This modification would mitigate their major weakness (i.e., the introduction of bias) but still will not be equivalent with the reproduction of certain, pre-specified, distribution functions. On top of this, since the ICDF is employed, a unique advantage of SMARTA (and other Nataf-based approaches) over the aforementioned schemes is that it can be used for the simulation of both univariate and multivariate stationary processes with discrete, continuous and mixed-type distributions. Regarding parameterization, the proposed Nataf-based approach exhibit a parsimonious character, as it is evident by the small number of required parameters, which are equal or lower than those required by the aforementioned schemes (for a comparison see section 4.1). Finally, it is noted that, due to the definition and use of Pearson's correlation coefficient (see Eq. (12)), none of the latter methods (including SMARTA), can be used for the simulation of processes characterized by distributions functions exhibiting infinite variance. In such situations the use of alternative simulation methods is required (e.g., Samoradnitsky, 2017). Random variables with infinite moments typically arise when heavy-tailed distribution functions with power-type tails are employed. For instance, a Pareto type-I distribution with CDF, $F(x) = 1 - (x/b)^{-a}$, where $b > 0$ (scale), $a > 0$ (shape) and $x \geq b$, has finite variance only for $a > 2$. The literature offers a plethora of studies indicating the suitability of heavy-tailed distributions for both precipitation (e.g., Cavanaugh et al., 2015; Koutsoyiannis & Papalexiou, 2016; Papalexiou et al., 2013; Papalexiou & Koutsoyiannis, 2013, 2016) and streamflow (e.g., Anderson & Meerschaert, 1998; Basso et al., 2015; Blum et al., 2017; Bowers et al., 2012) processes, especially regarding the description of their extreme behavior. After reviewing the outcomes of these studies, which involve the analysis of numerous worldwide historical records, we found that the majority of them, agree that the hydrological variables are characterized by distribution functions (with either exponential or power-type tails) with finite variance. On top of the empirical evidence provided by the aforementioned works, theoretical reasoning (related with entropy and energy production) further supports the finite variance hypothesis for hydrometeorological processes (Koutsoyiannis, 2016, 2017). In this vein, it is regarded that the finite variance assumption poses a practical barrier of limited impact, if any, on the application of latter methods for the simulation of hydrometeorological processes.

4 Hypothetical simulation studies

Prior to employing real-world datasets to demonstrate the proposed approach, we decided to setup two hypothetical simulation studies. One univariate and one multivariate. The motivation behind this choice was based on conducting experiments where all the assumptions are *a priori* known, hence allowing the comprehensive evaluation and assessment of the model without the effect of exogenous factors, such as, erroneous or short length historical data. However, it is remarked that the proposed method is generic, and can be directly applied for the simulation of univariate and multivariate stationary processes (e.g., geophysical, hydrometeorological and beyond). In that respect, in section 5 the applicability of SMARTA is demonstrated using two real-world datasets, one that concerns the simulation of annual non-Gaussian streamflow at four stations and another that involves the simulation of intermittent, non-Gaussian, daily rainfall at a single location.

4.1 Simulation of univariate processes

The first simulation study constitutes a comparison between the original SMA and the proposed SMARTA models (with $q = 2^{12}$ for both) for the simulation of long (i.e., 2^{20} time steps) univariate HK processes (i.e., fGn), exhibiting different Hurst coefficients, i.e., $H \in \{0.6, 0.7, 0.8, 0.9\}$ and Pearson type-III marginal distribution ($\mathcal{P}$III). With this in mind, we identified a total of 4 scenarios, each one characterized by $\mathcal{P}$III and different H coefficients. It is reminded that the original SMA model, in order to approximate the marginal statistics, uses $\mathcal{P}$III variates for the innovation term (hence hereafter referred to as SMA-$\mathcal{P}$III), while SMARTA uses the ICDF of the target distribution—in this case $\mathcal{P}$III. The rationale regarding the selection of this distribution was the intention to conduct a fair and meaningful comparison among the two models, which, in this formulation, have exactly the same number of parameters, i.e., three for the marginal distribution (see, Eq. (29)) and one (i.e., H) for the autocorrelation structure. We point out that, the comparison is not intended to infer which model is the best, but rather used as a benchmark to highlight the merits of the proposed approach. $\mathcal{P}$III is essentially a Gamma distribution (see, Eq. (16))) with an additional location (else known as threshold or shift) parameter, whose PDF is given by,

$$f_{\mathcal{P}\mathrm{III}}(x; a, b, c) = \frac{1}{|\mathrm{b}|\,\Gamma(a)}\left(\frac{x-c}{b}\right)^{a-1} \exp\left(-\frac{x-c}{b}\right), \begin{cases} \text{if } b > 0 & c \leq x < \infty \\ \text{if } b < 0 & -\infty < x \leq c \end{cases} \tag{29}$$

where $\Gamma(\cdot)$ denotes the gamma function, while, $a > 0$, $b \neq 0$ and $c \in \mathbb{R}$ are shape, scale and location parameters, respectively; and they are interconnected with the mean ($\mu_{\underline{x}}$), variance ($\sigma_{\underline{x}}^2$), skewness ($C_{s_{\underline{x}}}$) and kurtosis ($C_{k_{\underline{x}}}$) coefficients of random variable $\underline{x}$ by,

$$\mu_{\underline{x}} = c + ab, \qquad \sigma_{\underline{x}}^2 = ab^2, \qquad C_{s_{\underline{x}}} = \frac{2b}{|b|\sqrt{a}}, \qquad C_{k_{\underline{x}}} = \frac{6}{a} + 3 \tag{30}$$

More specifically, in all scenarios, we employed a $\mathcal{P}$III distribution with parameters $a = 0.75614, b = 11.5$ and $c = 1.30434$, whose theoretical moments are presented in **Table 2**.

Table 2. Summary of theoretical and simulated statistics as reproduced by SMA and SMARTA models.

	Theoretical	Simulated (SMA-$\mathcal{P}$III)				Simulated (SMARTA)			
Scenario	All	H=0.6	H = 0.7	H = 0.8	H = 0.9	H=0.6	H = 0.7	H = 0.8	H = 0.9
Mean (μ)	10	9.99	10.08	9.85	10.23	10.00	9.99	9.99	10.00
Variance (σ^2)	100	100.61	100.78	100.04	99.79	100.03	99.86	100.07	101.65
Skewness coeff. (C_s)	2.30	2.35	2.34	2.32	2.35	2.30	2.29	2.30	2.35
Kurtosis coeff. (C_k)	10.93	11.43	11.80	12.62	15.97	10.94	10.85	11.00	11.53
Hurst coeff. (H)	0.60, 0.70, 0.80, 0.90	0.61	0.70	0.80	0.89	0.60	0.71	0.80	0.90

*The theoretical moments correspond to $\mathcal{P}$III distribution (a = 0.75614, b = 11.5 and c = 1.30434).

Regarding SMARTA and the given marginal distribution, **Figure 4**a illustrates the relationship between the equivalent correlation coefficients $\tilde{\rho}$ and the target ones ρ (the superscripts are omitted for simplicity), while **Figure 4**b depicts the equivalent autocorrelation coefficients $\tilde{\rho}_\tau$ employed by SMARTA, in order to capture the target autocorrelation structure ρ_τ of the target HK processes.

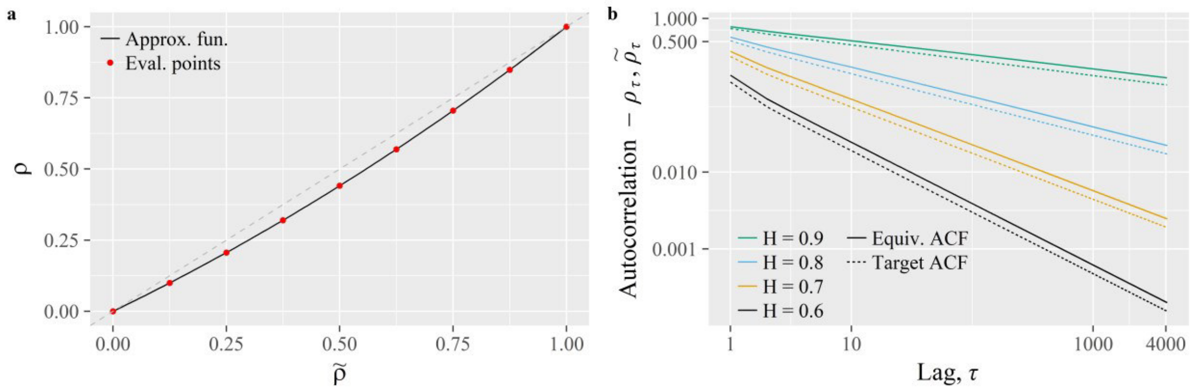

Figure 4. a) The established relationship between equivalent, $\tilde{\rho}$ and target ρ correlation coefficients. b) Comparison between the target and equivalent autocorrelation coefficients employed within the SMARTA model for HK processes with the various values of H.

Table 2 presents the simulated (by the two approaches) first four moments; which are apparently well-captured by both models. It is noted that, while SMA does not explicitly accounts for the kurtosis coefficient, it is able to reproduce it in a satisfactory degree; especially when one considers the high uncertainty associated with its estimation (cf., Lombardo et al., 2014). Nevertheless, it is reminded that the resemblance of the moments does not imply the reproduction of the marginal distribution (Matalas & Wallis, 1976). This is clearly depicted in **Figure 5**a-d, where we compare the target theoretical cumulative distribution (CDF) with the empirically derived cumulative density functions (ECDFs) of the two models. In this case, only SMARTA was able to reproduce the target distribution, regardless of the value of H coefficient (its ECDF is almost indistinguishable from the theoretical one). On the other hand, the ECDF of SMA-$\mathcal{P}$III departs from the theoretical one for high values of H (e.g., see **Figure 5**d). Furthermore, SMARTA explicitly avoids the generation of negative values; since the target distribution ($\mathcal{P}$III) is positively bounded at $c = 1.30434$. A property of high importance in hydrology due to the (often) non-negative nature of such variables (e.g., streamflow and precipitation).

Regarding the resemblance of the auto-dependence structure of the processes, it is apparent from **Figure 5**e-h and **Figure 5**i-l that, both models were able to reproduce the theoretical HK ACFs as well as the corresponding climacograms, even for high values of H. The latter graphs also provide an empirical evidence of the theoretical consistency of both approaches. In addition, the Hurst coefficient of the synthetic realizations (see **Table 2**) was estimated using the climacogram-based, least squares variance (LSV) method (Tyralis & Koutsoyiannis, 2011) and are in agreement with the theoretical values.

Finally, in order to visually assess the form of the established dependencies, for both models and each HK process (i.e., scenario), we employ scatter plots of the lagged synthetic data for $\tau = 1$ (**Figure 5**m-p) and $\tau = 10$ (**Figure 5**q-t). It is observed that, despite the fact that both models reproduced the same autocorrelation coefficient for $\tau = 1$ and $\tau = 10$, they establish particularly different dependence patterns. This is attributed to the underlying assumption of SMARTA regarding the joint behavior of the process which is related to the Gaussian copula (expressed through the auxiliary Gaussian model).

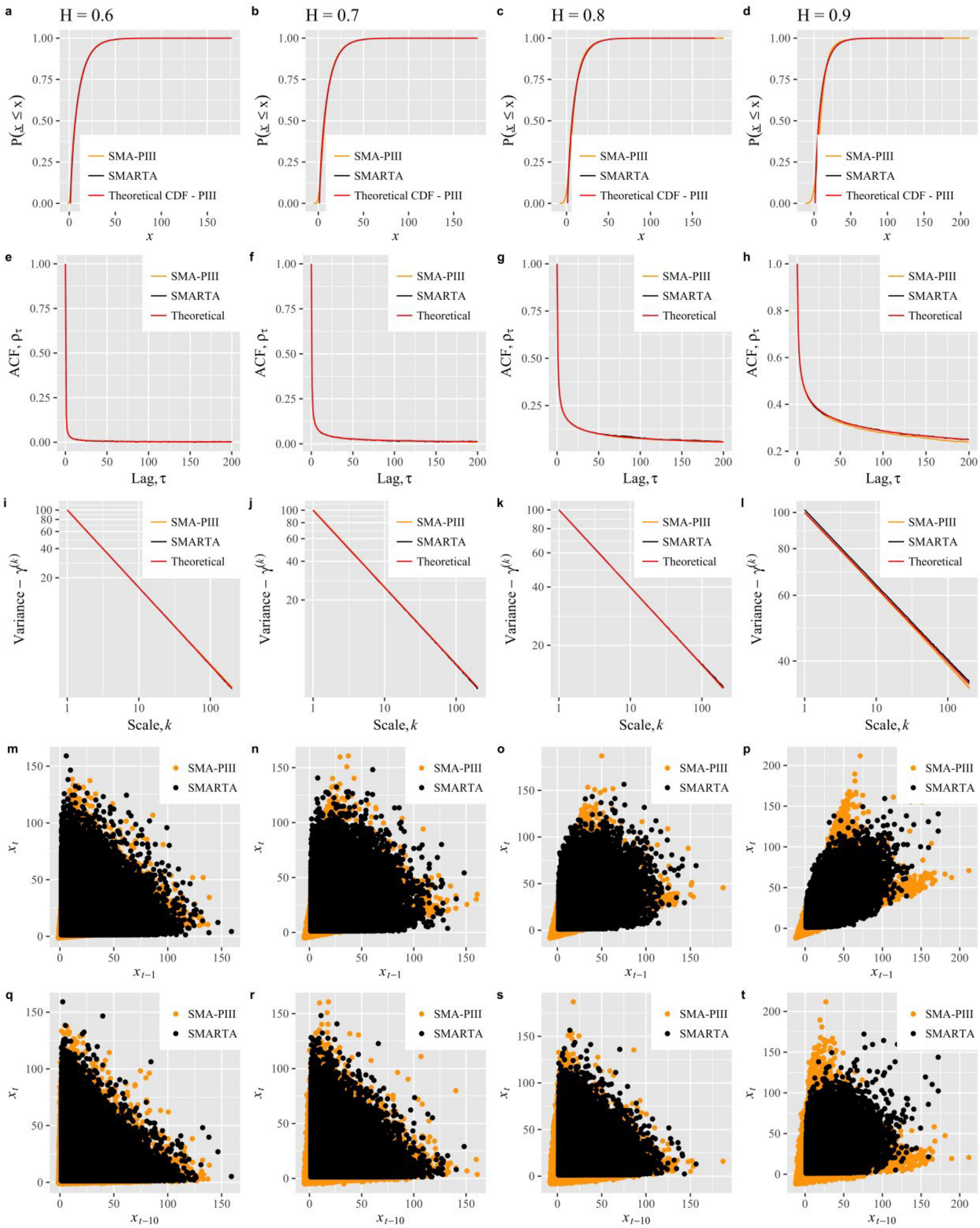

Figure 5. Comparison between theoretical and simulated CDFs (using the Weibull's plotting position) of SMA-$\mathcal{P}$III and SMARTA models for HK processes with a) $H = 0.6$, b) $H = 0.7$, c) $H = 0.8$, d) $H = 0.9$. Comparison between theoretical (HK) and empirical ACF of SMA-$\mathcal{P}$III and SMARTA models for HK processes with e) $H = 0.6$, f) $H = 0.7$, g) $H = 0.8$, h) $H = 0.9$. Comparison between theoretical (HK) and empirical climacograms of SMA-$\mathcal{P}$III and SMARTA models models for HK processes with i) $H = 0.6$, j) $H = 0.7$, k) $H = 0.8$, l) $H = 0.9$. Scatter plots of SMA-$\mathcal{P}$III and SMARTA models for time lag $\tau = 1$ for simulated HK processes with m) $H = 0.6$, n) $H = 0.7$, o) $H = 0.8$, p) $H = 0.9$. Scatter plots of SMA-$\mathcal{P}$III and SMARTA models for time lag $\tau = 10$ for simulated HK processes with q) $H = 0.6$, r) $H = 0.7$, s) $H = 0.8$, t) $H = 0.9$.

4.2 Simulation of multivariate processes

To further elaborate on the SMARTA approach, we setup a multivariate problem that concerns the simultaneous generation of four contemporaneously cross-correlated SRD and LRD processes. The latter may be seen as four (4) different processes at the same site, or processes of the same variable at 4 different sites. Hereinafter, we consider the latter for convenience and refer to them as sites A-D, as well as model them in that order, i.e., as 4-dimensional stationary process $\underline{\boldsymbol{x}}_t = \left[\underline{x}_t^1, \underline{x}_t^2, \underline{x}_t^3, \underline{x}_t^4\right]^{\mathrm{T}}$, where for instance, $i = 3$ refers to site C. In this demonstration, the target auto-dependence structure of each process is described by the two-parameter CAS (i.e., Eq. (6)). More specifically, sites A and B are characterized by LRD behavior (particularly HK, since we set $\beta > 1$ and $\kappa = \kappa_0$) and slowly-decaying ACF, while sites C and D by SRD (since we set $\beta = 0$) and fast-decaying ACF. In addition, we assigned different target distributions to the sites A-D, i.e., Burr type-XII (Eq. (31)), Pearson Type-III (Eq. (29)), Log-Normal (Eq. (17)) and Weibull (Eq. (32)). The PDF of the Burr type-XII distribution is given by,

$$f_{\mathcal{Br}\mathrm{XII}}(x; a_1, a_2, b) = \left(\frac{a_1 a_2}{b}\right)\left(\frac{x}{b}\right)^{a_1-1}\left(1+\left(\frac{x}{b}\right)^{a_1}\right)^{-a_2-1}, \qquad x > 0 \tag{31}$$

where $a_1, a_2 > 0$ are shape parameters and $b > 0$ is a scale parameter. It is noted that $\mathcal{Br}$XII is a power-type distribution and its r^{th} moment exist if and only if $a_1 a_2 > r$. Furthermore, the PDF of the Weibull reads as follows,

$$f_{\mathcal{WEI}}(x; a, b) = \left(\frac{a}{b}\right)\left(\frac{x}{b}\right)^{a-1}\exp\left(-\left(\frac{x}{b}\right)^{a}\right), \qquad x \geq 0 \tag{32}$$

where $a > 0$ and $b > 0$ are shape and scale parameters respectively. **Table 3**a provides a synopsis of the latter assumptions, as well as the parameters of CAS and the theoretical moments of the corresponding distributions. Note that, the Kurtosis coefficient of site A is infinite, since $a_1 a_2 < 4$. Further to this, the target and equivalent lag-0 cross-correlation coefficients (involving both positive and negative ones) are given in **Table 3**b. It is apparent that this is a peculiar simulation scenario, which was devised in order stress-test the SMARTA method.

Table 3. a) Synopsis of theoretical distribution models and their moments, as well as, of CAS parameters for each variable of the multivariate simulation study. b) The upper triangle (grey cells) contains the target lag-0 cross-correlation coefficients ($\rho_0^{i,j}$) between sites A-D, while the lower triangle depicts the corresponding estimated equivalent correlation coefficients ($\tilde{\rho}_0^{i,j}$).

a)

Distribution/ Parameters	Theoretical			
	Site A	Site B	Site C	Site D
	$\mathcal{Br}$XII	$\mathcal{P}$III	$\mathcal{LN}$	$\mathcal{WEI}$
a	2.5 (a_1)	3	0.5	1.5
b	1	1	2	10
c	1.5 (a_2)	10	-	-
Statistic	**Theoretical**			
Mean (μ)	4.76	13	8.37	9.02
Variance (σ^2)	11.42	3	19.91	37.56
Skewness coeff. (C_s)	5.01	1.15	1.75	1.07
Kurtosis coeff. (C_k)	-	8	8.89	4.39
CAS parameter, β	1.25	1.66	0	0
CAS parameter, κ	11.32	5	0.5	0.2
Hurst coeff. (H)	0.6	0.7	0.5	0.5

b)

Lag-0 cross-correlation				
	Site A	Site B	Site C	Site D
Site A	1	-0.700	0.750	0.600
Site B	-0.940	1	-0.600	-0.700
Site C	0.862	-0.749	1	0.650
Site D	0.811	-0.923	0.707	1

*Distribution abbreviations: $\mathcal{Br}$XII: Burr type-XII (a_1 = shape, a_2 = shape, b = scale), $\mathcal{P}$III: Pearson type-III (a = shape, b = scale, c = location), $\mathcal{LN}$: Log-Normal (a = shape, b = scale), $\mathcal{WEI}$: Weibull (a = shape, b = scale).

In order to provide further insights regarding the theoretical consistency of the model, we generated 100 independent realizations with length 2^{11} time steps and set the number of SMARTA model's internal weight coefficients equal to $q = 2^{10}$. **Figure 6** provides a synopsis of some basic dependence-related statistics in terms of box-plots. Clearly, SMARTA resembled with high precision the lag-1 autocorrelation and lag-0 cross-correlation coefficients (including the negative ones), despite the fact that the target processes are characterized by very different auto-dependence structures and distribution functions. Additionally, regarding the Hurst coefficient of the simulated series, it was once again estimated with the LSV method. A small discrepancy that concern site D, which is an SRD process (i.e., $H = 0.5$) is observed. This may be attributed to the associated estimation method and the high lag-1 autocorrelation (~0.8) of site D. Furthermore, in **Figure 7**a-d we compared the empirical distribution of each realization of each site A-D, with the corresponding theoretical distribution, in terms of the survival function (SF), also known as complementary CDF or tail function. The latter is denoted by $\overline{F}_{\underline{x}}$ and expresses the probability of exceedance, i.e., $\overline{F}_{\underline{x}} \coloneqq P(\underline{x} > x) = 1 - F_{\underline{x}}$. The latter figure highlights the ability of the model to preserve the target distribution functions, even in multivariate mode, since the median SF of all 100 realizations for the 4 sites is virtually identical to the associated theoretical model. Furthermore, in **Figure 7**e-h we depict the relationship between the equivalent, $\tilde{\rho}$ and target ρ correlation coefficients for each site A-D, while the preservation of the theoretical auto-dependence structure can be verified by the simulated ACFs (**Figure 7**i-l) and climacograms (**Figure 7**m-p) of the four variables, that closely resemble the corresponding theoretical ones. To further explore the joint behavior of the model and the established dependence patterns, we employ scatter plots. **Figure S1** of supplementary material (SM) depicts the established dependence patterns among the variables for time lag 0 (SM, **Figure S1**e, i, j, m, n, o), as well as for each variable for time lag 1 (SM, **Figure S1**a, f, k, p). Finally, the relationship between equivalent, $\tilde{\rho}^{i,j}$ and target $\rho^{i,j}$, correlation coefficients is provided for every combination of sites A-D (SM, **Figure S1**b, c, d, g, h, l).

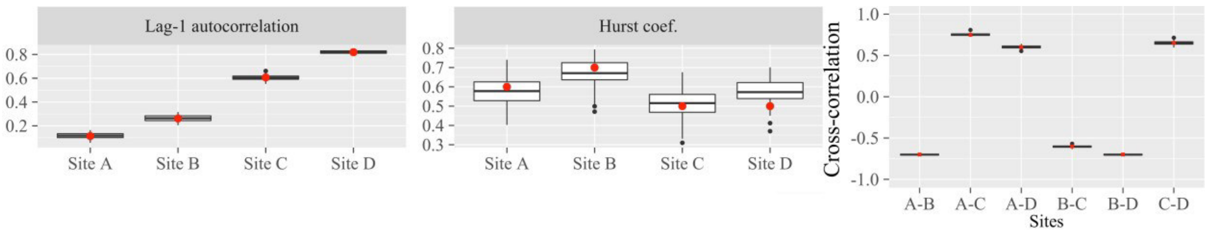

Figure 6. Comparison between theoretical (red dots, •) and simulated lag-1 autocorrelation and Hurst coefficient for sites A-D. Target (red dots, •) and simulated lag-0 cross-correlation coefficients for all pairs of sites A-D.

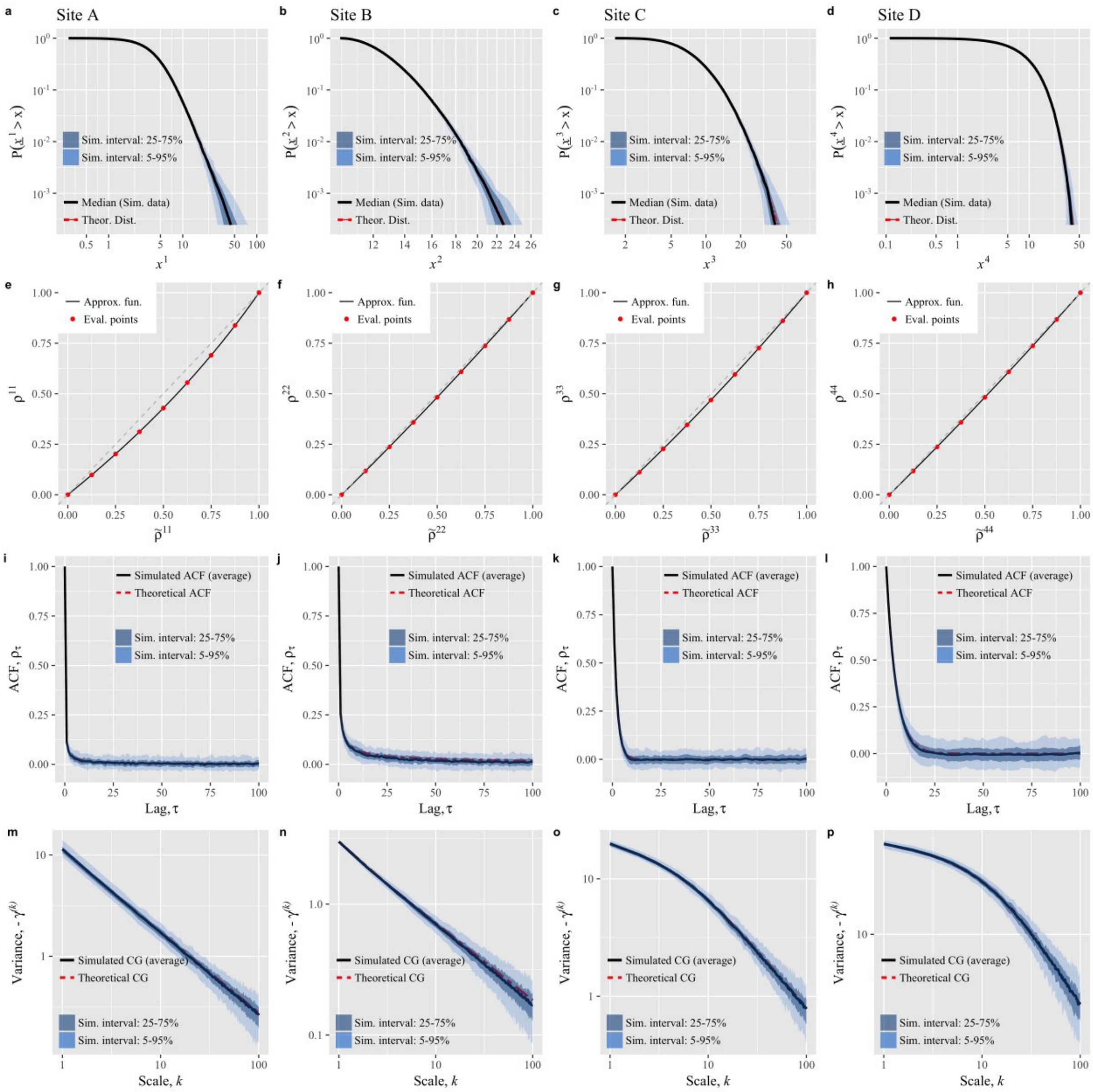

Figure 7. (a-d) Theoretical and simulated (SMARTA) distribution functions (using the Weibull's plotting position) for sites A-D. (e-h) The established relationships between equivalent, $\tilde{\rho}$ and target ρ correlation coefficients given the marginal distribution of sites A-D. (i-l) Theoretical and simulated ACFs for sites A-D. (m-p) Theoretical and simulated climacograms (CGs) for sites A-D. In all cases, the simulation intervals have been established using all 100 realizations.

5 Real-world simulation studies

5.1 Generation of multivariate annual streamflow time series

The first real-world simulation study concerns the application of SMARTA for the synthesis of annual streamflow time series at 4 stations in New South Wales region, Australia (Australian Government Bureau of Meteorology, 2015). Particularly, we employed historical data (**Figure 8**a-d) from the following stations: Maragle Creek at Maragle (ID1: 401009), Goobarragandra River at Lacmalac (ID2: 410057), Adelong Creek at Batlow Road (ID3: 410061), Cotter River at Gingera (ID4: 410730). Hereinafter, we refer to them using their station ID, as well as model them in that order, as 4-dimensional stationary process $\underline{\boldsymbol{x}}_t = [\underline{x}_t^1, \underline{x}_t^2, \underline{x}_t^3, \underline{x}_t^4]^{\mathrm{T}}$; (i.e., $i = 3$ refers to station Adelong Creek at Batlow Road with ID3: 410061). The distribution of historical data does not exhibit the typical bell-type shape that is often encountered in annual data, hence we use the Gamma and Weibull distributions to model them. Specifically, using the maximum likelihood estimation method we identified the following distributions, $\underline{x}_t^1 \sim \mathcal{G}(a = 2.13, b = 16.95)$, $\underline{x}_t^2 \sim \mathcal{WEI}(a = 2.30, b = 302.11)$, $\underline{x}_t^3 \sim \mathcal{WEI}(a = 2.40, b = 15.75)$ and $\underline{x}_t^4 \sim \mathcal{G}(a = 1.95, b = 48.48)$. In addition, they are characterized by moderate-to-high temporal dependence and high lag-0 cross-correlation coefficients, that range from 0.83 ($\rho_0^{1,4}$) to 0.93 ($\rho_0^{2,3}$). Following Koutsoyiannis (2000), the parameters of CAS (i.e., Eq. (6) - given in vector format), $\boldsymbol{\beta} = [0.99, 0.75, 1.13, 0.72]$ and $\boldsymbol{\kappa} = [2.57, 4.41, 6.01, 5.07]$ were identified for each process by minimizing the mean square error (MSE) among the sample and theoretical autocorrelation coefficients. In this case study, we simulated one realization of 1 000 years using the SMARTA model (with $q = 2^9$). **Figure 8**e-h provides, for each station, a visual comparison among the empirical, theoretical and simulated distribution. Furthermore, **Figure 8**i-l depicts, for each process, the relationship between the equivalent and target autocorrelation coefficients. The ability of the model to establish the target auto-dependence structures is verified by comparing, the theoretical and simulated ACF (**Figure 8**m-p) and corresponding climacogram (**Figure 8**q-t) of each process. Finally, the model reproduced the target lag-0 cross-coefficients with high accuracy (SM, **Figure S2**) and established dependence patterns that are in agreement with the observed ones (SM, **Figure S2**).

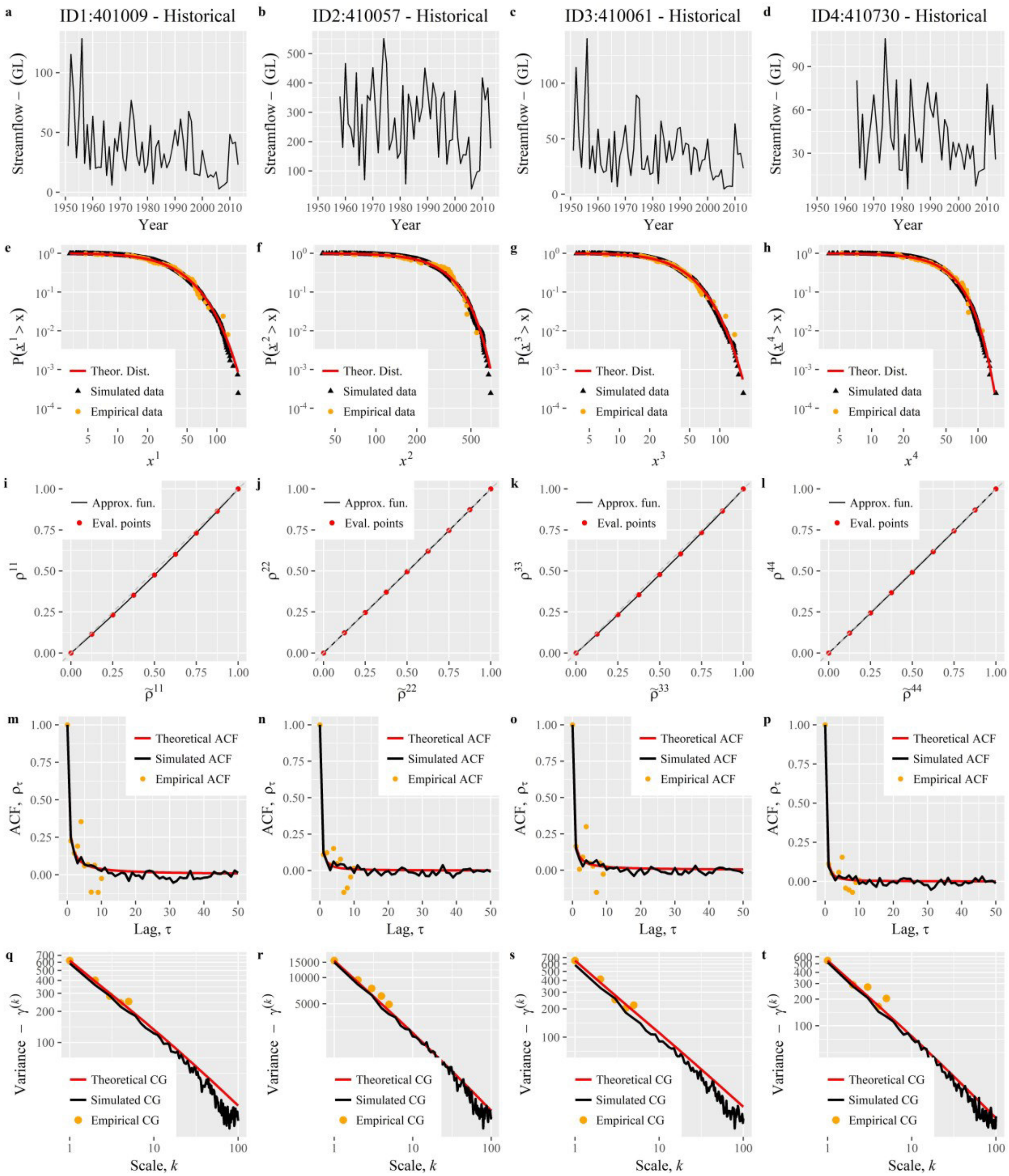

Figure 8. Synopsis of annual streamflow simulation study at 4 stations in New South Wales region. (a-d) Historical time series. (e-h) Empirical, simulated and theoretical distribution functions (using the Weibull's plotting position) for stations ID1-4 (i-l) The established relationships between equivalent, $\tilde{\rho}$ and target ρ correlation coefficients given the marginal distribution of stations ID1-4. (m-p) Empirical, simulated and theoretical ACFs for stations ID1-4. (q-t) Empirical, simulated and theoretical climacograms (CGs) for stations ID1-4.

5.2 Generation of univariate daily rainfall time series

In the final case study, we employ SMARTA for the stochastic simulation of a univariate daily rainfall process characterized by intermittency. The available data concern an observation period spanning from 1/1/1964 to 31/12/2006 (43 years) from Pavlos rain gauge located at Boeticos Kephisos river basin, Greece (**Figure 9**a). See also Efstratiadis et al. (2014) for further details regarding the dataset. In general, apart from *ad-hoc* techniques to handle intermittency (e.g., truncation to zero of values below a threshold), typical stochastic simulation schemes (e.g., Papalexiou, 2018; Serinaldi, 2009; Serinaldi & Kilsby, 2014) rely on the use of mixed-distributions or employ two-part models, which, in a nutshell, describe precipitation processes as the product of two different processes, particularly, that of occurrence (rain or no-rain) and that of intensity (e.g., Ailliot et al., 2015; Breinl et al., 2013; Brissette et al., 2007; Khalili et al., 2009; Lee, 2016, 2017; Lombardo et al., 2017; Mhanna & Bauwens, 2012; Thompson et al., 2007; Wilks, 1998; Wilks & Wilby, 1999). Herein, we employ the former approach, that is, mixed-distributions, as it seems a convenient option (Papalexiou, 2018) given the characteristics of SMARTA and particularly its flexibility regarding the selection of the marginal distribution. An alternative option, also compatible with the proposed method (and Nataf-based schemes in general), would be the use of single distribution functions that exhibit an atom of probability mass at zero. A characteristic example, which in the past has been used for this purpose (Dunn, 2004; Hasan & Dunn, 2011), is the Tweedie distribution (Jorgensen, 1987; Tweedie, 1984). Nevertheless, in this simulation study, in order to simultaneously account for the effect of seasonality and the stationarity assumption of the model, we treat each month as separate stochastic process, by varying the distribution function and autocorrelation structure on a monthly basis. Specifically, regarding the marginal distribution, we employ a discrete–continuous (i.e., mixed or zero-inflated) model whose CDF is given by,

$$F_{\underline{x}}(x) = \begin{cases} p_D, & x \leq 0 \\ p_D + (1 - p_D) G_{\underline{x}}(x), & x > 0 \end{cases} \tag{33}$$

where, p_D denotes the probability of a dry interval (abbreviated as probability dry), i.e., $p_D \coloneqq P(\underline{x} \leq x_D)$ and $G_{\underline{x}}$ stands for the distribution of amounts greater than the threshold x_D, i.e., $G_{\underline{x}} \coloneqq F_{\underline{x}|\underline{x}>x_D} = P(\underline{x}|\underline{x} > x_D)$. Moreover, the corresponding ICDF is given by,

$$F_{\underline{x}}^{-1}(u) = \begin{cases} 0, & 0 \leq u \leq p_D \\ G_{\underline{x}}^{-1}\left(\dfrac{(u - p_D)}{(1 - p_D)}\right), & p_D < u \leq 1 \end{cases} \tag{34}$$

where $u \in [0, 1]$ denotes probability. In this formulation values less or equal to x_D (that arise with probability p_D) are assumed equal to zero. We remind the reader that the solely requirement of the algorithm of Appendix A, that is used to approximate the relationship $\mathcal{F}(\cdot)$ of Eq. (15), hence the equivalent correlations $\tilde{\rho}_\tau$, is the ICDF (thus conveniently accounting for mixed distributions; e.g., Eq. (33)). Nevertheless, after the specification of the threshold x_D, the empirical probability dry, p_D, can be directly obtained from the available data by counting the number of dry occurrences and dividing it with the total number of observed data. Regarding, $G_{\underline{x}}$, it is obtained by selecting and fitting a theoretical distribution to the amount data above threshold x_D. In this demonstration, we set $x_D \coloneqq 0$, and for the description of the positive daily precipitation amounts of all months, we employ the generalized gamma ($\mathcal{GG}$) distribution (Stacy, 1962), which has been proved particularly capable for the task at hand (Chen et al., 2017; Papalexiou, 2018; Papalexiou & Koutsoyiannis, 2016). Of course, depending on the case, the $\mathcal{GG}$ could be replaced with other distribution functions. Back in our case, the parameters of the $\mathcal{GG}$ distribution were identified using a fitting approach based on L-moments (Hosking, 1990); specifically the one proposed by Papalexiou and Koutsoyiannis (2016). The PDF of $\mathcal{GG}$ distribution is given by,

$$f_{\mathcal{GG}}(x; a_1, a_2, b) = \frac{a_2}{b\Gamma(a_1/a_2)} \left(\frac{x}{b}\right)^{a_1 - 1} \exp\left(-\left(\frac{x}{b}\right)^{a_2}\right), \qquad x > 0 \tag{35}$$

where $\Gamma(\cdot)$ denotes the gamma function, while, $a_1 > 0, a_2 > 0$ are parameters that control the shape of the distribution and $b > 0$ is a scale parameter. The interested reader is referred to the latter works for further details regarding the $\mathcal{GG}$ distribution and the associated fitting method.

For instance, concerning the marginal characteristics of October's daily rainfall, we estimated the probability dry, $p_D = 0.84$, while the parameters of $\mathcal{GG}$ were found $b = 3.96$, $a_1 = 0.851$ and $a_2 = 0.588$. Furthermore, regarding the description of the auto-dependence structure of the process, we employed CAS and estimated its parameters on a monthly basis (e.g., for October it we identified, $\beta = 0$ and $\kappa = 1.36$) by minimizing the MSE among the sample and theoretical autocorrelation coefficients. Finally, we generated 1 000 years (i.e., 365 000 days) of synthetic data (**Figure 9**b depicts a random window of 60 years) and performed a similar analysis with the previous cases studies; which is summarized in **Figure 9**, where we depict the results of three characteristic months, i.e., February, June and October (the results are similar for the other months – see SM, **Figure S3**-**S6**). Particularly, panels (c)-(e) illustrate the capability of the model to reproduce the target distributions (in terms of the SF) of positive precipitation amounts (p_D is explicitly preserved since it is embedded in the employed mixed-distribution model), while, panels (f)-(h) depicts the relationship of equivalent, $\tilde{\rho}$ and target ρ correlation coefficients for both $\mathcal{GG}$ and mixed-distribution models. It is observed that, the non-linearity of this relationship increases from $\mathcal{GG}$ to mixed distribution due to the fact that the latter is zero-inflated. Furthermore, panels (i)-(k) depict the accurate resemblance of the target autocorrelation structure (i.e., CAS), while, panels (l)-(n) provide a comparison of empirical and simulated scatter for time lag 1, which seems to be in agreement with the historical pattern. Finally, preliminary analysis (not shown herein) indicated that the model has the potential to approximate some of the empirical statistics (in terms of L-moments) across coarser time scales, even though they are not explicitly modelled by it. This observation should not be interpreted as a general conclusion, rather as a direction for further investigation. We remark that the literature offers several well-established techniques with proven results, specifically designed for this purpose, i.e., to address scaling and intermittency, such as disaggregation (e.g., Kossieris et al., 2016; Lombardo et al., 2017) and multi-fractal methods, based on cascade models (e.g., Deidda et al., 1999; Kantelhardt et al., 2006; Tessier et al., 1996). The latter methods, by design, aim to simultaneously resemble the process at multiple aggregation levels, employing scaling relationships for high order moments (often greater than second). In our view, an interesting topic of future research would be a comparison among the latter simulation techniques with Nataf-based methods for the reproduction of the multi-scale behavior that characterizes hydrometeorological processes. Similar works, yet involving alternative simulation schemes, are those of Lombardo et al. (2012) and Pui et al. (2012).

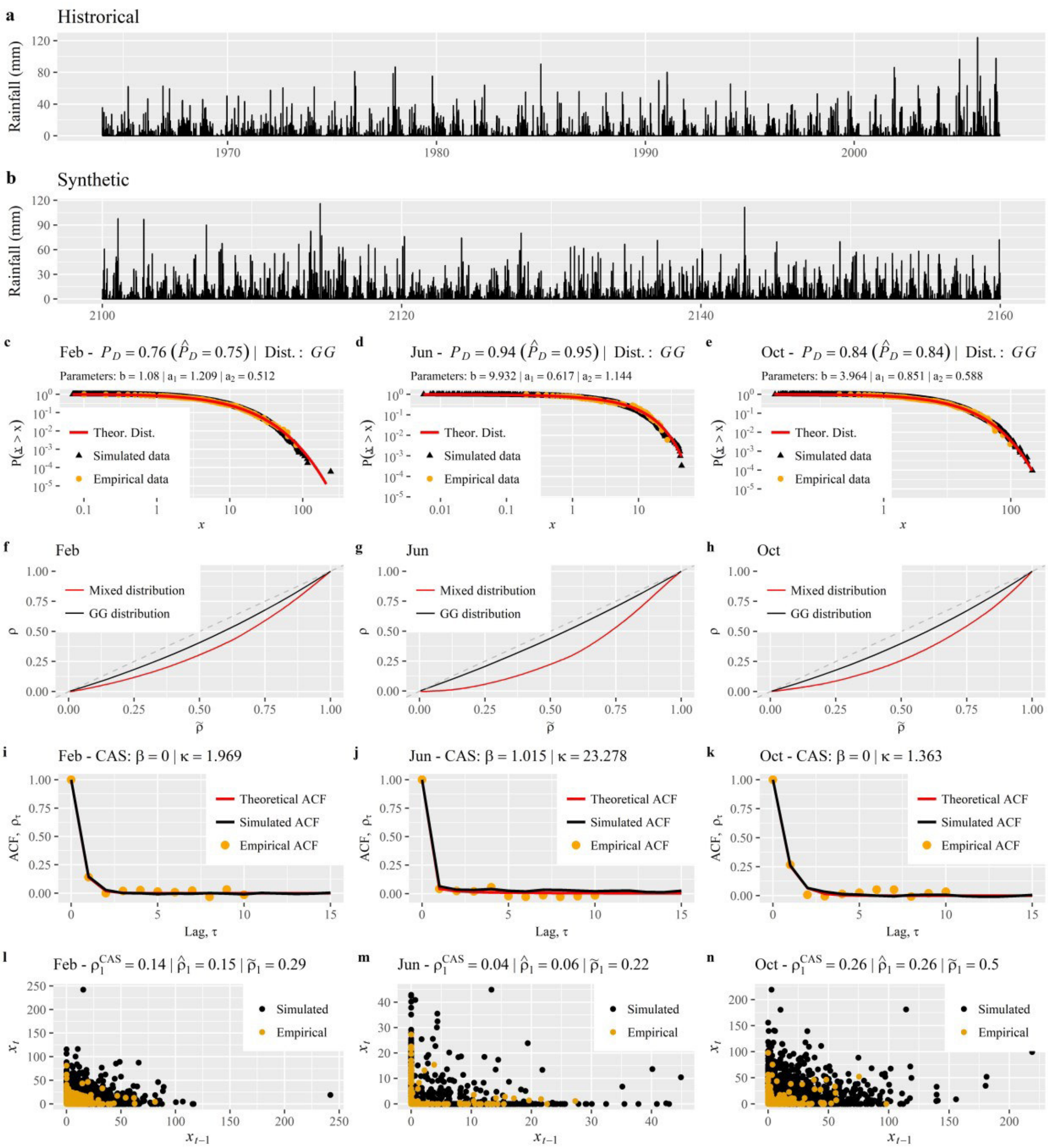

Figure 9. Synopsis of daily rainfall simulation at Pavlos' station. a) Historical time series. b) Synthetic time series; randomly selected window of 60 years. Empirical, simulated and theoretical distribution function of positive precipitation amounts for c) February, d) June and e) October (using the Weibull's plotting position); the title of each plot provides the parameters of the $\mathcal{GG}$ distribution, as well as the historical (p_D) and simulated ($\hat{p}_D$) values of probability dry. The established relationship between equivalent, $\tilde{\rho}$ and target ρ correlation coefficients for the mixed and $\mathcal{GG}$ distribution for f) February, g) June and h) October. Empirical, simulated and theoretical ACF for i) February, j) June and k) October; the title of each plot depicts the parameters of CAS. Empirical and simulated dependence pattern for time lag 1 for l) February, m) June and n) October; the title of each plot depicts the lag-1, target (ρ_1^{CAS}), simulated ($\hat{\rho}_1$), and equivalent ($\tilde{\rho}_1$) autocorrelation coefficients.

6 Conclusions

This paper introduces a novel and versatile stochastic model, termed SMARTA, with solid theoretical background and proven capability of addressing important hydrometeorological simulation problems. A prominent characteristic of the model is its ability to simulate univariate and multivariate stationary processes with any autocorrelation structure and marginal distribution, provided that the former is feasible and the latter have finite variance. The central idea of the method relies on the use of an appropriately parameterized (expressed through *equivalent* correlation coefficients) auxiliary Gaussian process which after its mapping to the actual domain results in a process with the desired stochastic structure and marginal distribution.

Briefly, the proposed approach is built upon three major elements: a) The SMA scheme of Koutsoyiannis (2000), which is used as an auxiliary model in the Gaussian domain, b) a generalized autocorrelation structure, that allows the parsimonious description of SRD and LRD processes, and c) the rationale of NDM (Nataf, 1962), and the associated mapping procedure, that provide the theoretical basis of the method and in turn allows the identification of the *equivalent* correlation coefficients; hence determine the parameters of the auxiliary model.

Overall, the proposed methodology maintains the flexible and parsimonious character of the original SMA model and simultaneously exhibit a series of additional virtues, as demonstrated through two hypothetical and two real-world simulation studies. Among them:

a) The unambiguous advantage of explicitly simulating any-range dependent (SRD or LRD) stationary processes with arbitrary distributions (even from different families, see section 4.2), using a single simulation scheme.
b) Its ability to simulate univariate and multivariate processes that exhibit contemporaneous cross-correlations. The generation of time series at multiple locations, or of individual correlated processes, is often the case in hydrological studies, making SMARTA a useful method for such tasks.
c) The possible incorporation of novel advances in statistical science in stochastic simulation; such as new distributions and robust fitting methods (e.g., L-moments). In addition, regarding distributions of hydrometeorological processes, SMARTA can take advantage of years of research in statistical analysis of hydrometeorological variables, since it can incorporate any distribution function whose variance exists.
d) The ability of the model to explicitly avoid the generation of negative values, which simultaneously is a shortcoming of many linear stochastic models. This is due to the direct use of the distribution function(s) within the generation mechanism of the model. If the latter is defined in the positive real line, then all the generated values will be within those bounds (i.e., positive).

Typical, but not limited, applications of SMARTA entail the simulation of stationary processes at time scales not affected by cyclostationary correlation structures (e.g., monthly scale). For instance, given the wide range of admissible correlation structures and distributions, it could be applied for the generation of synthetic time series at annual and fine time scales (e.g., daily) for various hydrometeorological processes, such as, precipitation, streamflow and temperature. The latter time series can be used as input in a variety of water resources risk-related studies and it is anticipated to improve the quality of their outcomes, due to more accurate representation of the input processes. Ongoing research aims in an enhanced stochastic simulation scheme that will combine (using disaggregation techniques) both stationary (e.g., SMARTA) and cyclostationary Nataf-based models (Tsoukalas et al., 2017, 2018a), thus providing an even more flexible and versatile simulation method for synthetic time series generation.

Acknowledgments

We would like to thank the Associate Editor and the three anonymous reviewers, for their constructive comments and critique, which helped providing a much-improved manuscript. Furthermore, we would like to sincerely thank P. Kossieris and G. Moraitis for their valuable comments and suggestions, as well as the many fruitful discussions. **Data availability**: The Australian streamflow data were retrieved from the Australian Bureau of Meteorology and they are available at: www.bom.gov.au/water/hrs/. The historical dataset of precipitation gauge Pavlos, Boeticos Kephisos river basin, Greece, can be obtained from the following link: http://kyy.hydroscope.gr/timeseries/d/216/. The aforementioned datasets can be also found in: http://www.itia.ntua.gr/en/docinfo/1863/.

Appendix A

Tsoukalas et al. (2018a) proposed a generic, yet simple and efficient method for the establishment of the relationship of Eq. (15), that concerns the estimation of equivalent correlation coefficients $\tilde{\rho}_{\xi,\psi}$ required by Nataf-based schemes. The method is essentially a combination of Monte-Carlo simulation and polynomial approximation and is applicable for discrete, mixed and continuous-type marginal distributions; since its only requirement is the ICDF. The basic steps of the algorithm are synopsized below (the indices were omitted for simplicity):

Let $\underline{x}_\xi$ and $\underline{x}_\psi$ be two random variables while $\tilde{\rho}_{\xi,\psi}$ and $\rho_{\xi,\psi}$ stand for the equivalent (in Gaussian domain) and the target correlation coefficients respectively. Furthermore, let $F_{\underline{x}_\xi}$ and $F_{\underline{x}_\psi}$, denote the corresponding target distributions, whose variance is assumed finite.

Step 1. Create a Ω-dimensional, equally spaced, vector $\tilde{\boldsymbol{r}} = \left[\tilde{r}^1, \dots, \tilde{r}^i, \dots, \tilde{r}^\Omega\right]$ in the interval $[r_{min}, r_{max}]$. Lemma 2 (see section 3.2) can be employed in order to determine the values of r_{min} and r_{max} since it provides insights regarding the sign of $\tilde{\rho}_{\underline{\xi},\underline{\psi}}$. For instance, if the target correlation $\rho_{\underline{\xi},\underline{\psi}}$ is positive we restrict our attention on the interval [0, 1].

Step 2. For each value of $\tilde{\boldsymbol{r}}$ generate N samples from the bivariate standard normal distribution with correlation $\tilde{r}^i$.

Step 3. Map the generated values to actual domain using their ICDF (i.e., $F_{\underline{x}_\xi}$ and $F_{\underline{x}_\psi}$) as in Eq. (10).

Step 4. Calculate and store the resulting correlation r^i in the vector $\boldsymbol{r} = \left[r^1, \dots, r^i, \dots, r^\Omega\right]$.

Step 5. Since Eq. (15) is a continuous function, bounded in the interval $[r_{min}, r_{max}]$, according to Weierstrass approximation theorem it can be approximated by a p-order polynomial of the form of Eq. (A.1) between $\tilde{r}$ and r.

$$\rho = \mathcal{F}\left(\tilde{\rho}\middle|F_{\underline{\xi}}, F_{\underline{\psi}}\right) \approx r = a_p\tilde{r}^p + a_{p-1}\tilde{r}^{p-1} + \dots + a_1\tilde{r}^1 + a_0 \tag{A.1}$$

Note that the constant term a_0 could be omitted as indicated by Lemma 2. Furthermore, in order to avoid over-fitting and possible ill-conditions, which could lead to simulation errors, the order of the polynomial can be determined with the use of cross-validation or Akaike information criterion (AIC). Alternatively, the degrees of freedom of the polynomial can be restricted (as in Xiao (2014)) by setting $p = \Omega - 1$. The latter author, based on a systematic analysis of a variety of distributions characterized by wide combinations of skewness and kurtosis coefficients, argued that the relationship of Eq. (15) can be well approximated by a polynomial of less than ninth degree ($p < 9$); hence proposed setting $\Omega = 9$ and $p = 8$. Moreover, it is noted that instead of a polynomial relationship, other type of functions can be used (e.g., Papalexiou, 2018; Serinaldi & Lombardo, 2017).

Step 6. Given a target correlation $\rho_{\underline{\xi},\underline{\psi}}$, evaluate the equivalent correlation $\tilde{\rho}_{\underline{\xi},\underline{\psi}}$ by inverting the fitted polynomial of Eq. (A.1).

It is remarked that the implementation of the latter algorithm in high-level programming languages (e.g., R or MATLAB) is fairly easy and straightforward, while a single run requires less than 0.5 second (with N = 150 000 and Ω = 9) on a typical 3.0 GHz Intel Dual-Core i5 processor with 4 GB RAM. Finally, it is noted that since it is a Monte-Carlo based method, the three parameters N, Ω and p control its accuracy and computational efficiency.

References

Ailliot, P., Allard, D., Monbet, V., & Naveau, P. (2015). Stochastic weather generators: an overview of weather type models. *Journal de La Société Française de Statistique*, *156*(1), 101–113.

Anderson, P. L., & Meerschaert, M. M. (1998). Modeling river flows with heavy tails. *Water Resources Research*, *34*(9), 2271–2280.

Australian Government Bureau of Meteorology. (2015). *Hydrologic reference stations, Bureau of Meteorology. [Available at: www.bom.gov.au/water/hrs/]*.

Basso, S., Schirmer, M., & Botter, G. (2015). On the emergence of heavy-tailed streamflow distributions. *Advances in Water Resources*, *82*, 98–105.

Beran, J. (1992). Statistical Methods for Data with Long-Range Dependence. *Statistical Science*, *7*(4), 404–416. https://doi.org/10.1214/ss/1177011122

Beran, J. (1994). *Statistics for long-memory processes* (Vol. 61). CRC press.

Biller, B., & Nelson, B. L. (2003). Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Transactions on Modeling and Computer Simulation*, *13*(3), 211–237. https://doi.org/10.1145/937332.937333

Blum, A. G., Archfield, S. A., & Vogel, R. M. (2017). On the probability distribution of daily streamflow in the United States. *Hydrology and Earth System Sciences*, *21*(6), 3093–3103. https://doi.org/10.5194/hess-21-3093-2017

Bowers, M. C., Tung, W. W., & Gao, J. B. (2012). On the distributions of seasonal river flows: Lognormal or power law? *Water Resources Research*, *48*(5), 1–12. https://doi.org/10.1029/2011WR011308

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bras, R. L., & Rodríguez-Iturbe, I. (1985). *Random functions and hydrology*. Addison-Wesley, Reading, Mass.

Breinl, K., Turkington, T., & Stowasser, M. (2013). Stochastic generation of multi-site daily precipitation for applications in risk management. *Journal of Hydrology*, *498*, 23–35. https://doi.org/10.1016/j.jhydrol.2013.06.015

Brissette, F. P., Khalili, M., & Leconte, R. (2007). Efficient stochastic generation of multi-site synthetic precipitation data. *Journal of Hydrology*, *345*(3–4), 121–133. https://doi.org/10.1016/j.jhydrol.2007.06.035

Camacho, F., McLeod, A. I., & Hipel, K. W. (1985). Contemporaneous autoregressive-moving average (CARMA) modeling in water resources. *Journal of the American Water Resources Association*, *21*(4), 709–720. https://doi.org/10.1111/j.1752-1688.1985.tb05384.x

Cario, M. C., & Nelson, B. L. (1996). Autoregressive to anything: Time-series input processes for simulation. *Operations Research Letters*, *19*(2), 51–58. https://doi.org/10.1016/0167-6377(96)00017-X

Cario, M. C., & Nelson, B. L. (1997). *Modeling and generating random vectors with arbitrary* *marginal distributions and correlation matrix*. *Industrial Engineering*. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.281&rep=rep1&type=pdf

Cavanaugh, N. R., Gershunov, A., Panorska, A. K., & Kozubowski, T. J. (2015). The probability distribution of intense daily precipitation. *Geophysical Research Letters*, *42*(5), 1560–1567.

Celeste, A. B., & Billib, M. (2009). Evaluation of stochastic reservoir operation optimization models. *Advances in Water Resources*, *32*(9), 1429–1443. https://doi.org/10.1016/j.advwatres.2009.06.008

Chen, H. (2001). Initialization for NORTA: Generation of Random Vectors with Specified Marginals and Correlations. *INFORMS Journal on Computing*, *13*(4), 312–331. https://doi.org/10.1287/ijoc.13.4.312.9736

Chen, L., Singh, V., & Xiong, F. (2017). An Entropy-Based Generalized Gamma Distribution for Flood Frequency Analysis. *Entropy*, *19*(12), 239. https://doi.org/10.3390/e19060239

Deidda, R., Benzi, R., & Siccardi, F. (1999). Multifractal modeling of anomalous scaling laws in rainfall. *Water Resources Research*, *35*(6), 1853–1867. https://doi.org/10.1029/1999WR900036

Demirtas, H., & Hedeker, D. (2011). A Practical Way for Computing Approximate Lower and Upper Correlation Bounds. *The American Statistician*, *65*(2), 104–109. https://doi.org/10.1198/tast.2011.10090

Der Kiureghian, A., & Liu, P.-L. (1986). Structural reliability under incomplete probability information. *Journal of Engineering Mechanics*, *112*(1), 85–104.

Detzel, D. H. M., & Mine, M. R. M. (2017). Comparison between Deseasonalized Models for Monthly Streamflow Generation in a Hurst?Kolmogorov Process Framework. *Journal of Hydrologic Engineering*, *22*(4), 05016040. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001488

Dimitriadis, P., & Koutsoyiannis, D. (2015). Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes. *Stochastic Environmental Research and Risk Assessment*, *29*(6), 1649–1669. https://doi.org/10.1007/s00477-015-1023-7

Ditlevsen, O. (1971). *Extremes and first passage times with applications in civil engineering*. Technical University of Denmark.

Dunn, P. K. (2004). Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, *24*(10), 1231–1239.

Efstratiadis, A., Dialynas, Y. G., Kozanis, S., & Koutsoyiannis, D. (2014). A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence. *Environmental Modelling and Software*, *62*(July), 139–152. https://doi.org/10.1016/j.envsoft.2014.08.017

Embrechts, P., McNeil, A. J., & Straumann, D. (1999). Correlation and Dependence in Risk Management: Properties and Pitfalls. In M. A. H. Dempster (Ed.), *Risk Management* (pp.

176–223). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511615337.008

Esscher, F. (1924). On a method of determining correlation from the ranks of the variates. *Scandinavian Actuarial Journal*, *1924*(1), 201–219.

Fatichi, S., Ivanov, V. Y., & Caporali, E. (2011). Simulation of future climate scenarios with a weather generator. *Advances in Water Resources*, *34*(4), 448–467. https://doi.org/10.1016/j.advwatres.2010.12.013

Feng, M., Liu, P., Guo, S., Gui, Z., Zhang, X., Zhang, W., & Xiong, L. (2017). Identifying changing patterns of reservoir operating rules under various inflow alteration scenarios. *Advances in Water Resources*, *104*, 23–36. https://doi.org/10.1016/j.advwatres.2017.03.003

Fowler, H. J., Kilsby, C. G., & O'Connell, P. E. (2000). A stochastic rainfall model for the assessment of regional water resource systems under changed climatic condition. *Hydrology and Earth System Sciences*, *4*(2), 263–281. https://doi.org/10.5194/hess-4-263-2000

Fréchet, M. (1957). Les tableaux de corrélation et les programmes linéaires. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *25*(1/3), 23. https://doi.org/10.2307/1401672

Giuliani, M., Herman, J. D., Castelletti, A., & Reed, P. (2014). Many-objective reservoir policy identification and refinement to reduce policy inertia and myopia in water management. *Water Resources Research*, *50*(4), 3355–3377. https://doi.org/10.1002/2013WR014700

Gneiting, T. (2000). Power-law correlations, related models for long-range dependence and their simulation. *Journal of Applied Probability*, *37*(4), 1104–1109.

Gneiting, T., & Schlather, M. (2004). Stochastic Models That Separate Fractal Dimension and the Hurst Effect. *SIAM Review*, *46*(2), 269–282. Retrieved from http://www.jstor.org/stable/20453506

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series and fractional differencing. *Journal of Time Series Analysis*, *1*(1), 15–29. https://doi.org/10.1111/j.1467-9892.1980.tb00297.x

Haberlandt, U., Hundecha, Y., Pahlow, M., & Schumann, A. H. (2011). Rainfall generators for application in flood studies. In *Flood Risk Assessment and Management* (pp. 117–147). Springer.

Hasan, M. M., & Dunn, P. K. (2011). Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia. *International Journal of Climatology*, *31*(9), 1389–1397.

Herman, J. D., Zeff, H. B., Lamontagne, J. R., Reed, P. M., & Characklis, G. W. (2016). Synthetic Drought Scenario Generation to Support Bottom-Up Water Supply Vulnerability Assessments. *Journal of Water Resources Planning and Management*, 04016050. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000701

Higham, N. J. (2002). Computing the nearest correlation matrix--a problem from finance. *IMA Journal of Numerical Analysis*, *22*(3), 329–343. https://doi.org/10.1093/imanum/22.3.329

Hoeffding, W. (1994). Scale—invariant correlation theory. In *The collected works of Wassily Hoeffding* (pp. 57–107). Springer.

Hosking, J. R. . M. (1984). Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research*, *20*(12), 1898–1908. https://doi.org/10.1029/WR020i012p01898

Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*(1), 105–124. Retrieved from http://www.jstor.org/stable/2345653

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.*, *116*, 770–808.

Iliopoulou, T., Papalexiou, S. M., Markonis, Y., & Koutsoyiannis, D. (2016). Revisiting long-range dependence in annual precipitation. *Journal of Hydrology*, *6*(4), 399–401. https://doi.org/10.1016/j.jhydrol.2016.04.015

Jorgensen, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *49*(2), 127–162. Retrieved from http://www.jstor.org/stable/2345415

Kantelhardt, J. W., Koscielny-Bunde, E., Rybski, D., Braun, P., Bunde, A., & Havlin, S. (2006). Long-term persistence and multifractality of precipitation and river runoff records. *Journal of Geophysical Research*, *111*(D1), D01106. https://doi.org/10.1029/2005JD005881

Kelly, K. S., & Krzysztofowicz, R. (1997). A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrology and Hydraulics*, *11*(1), 17–31. https://doi.org/10.1007/BF02428423

Khalili, M., Brissette, F., & Leconte, R. (2009). Stochastic multi-site generation of daily weather data. *Stochastic Environmental Research and Risk Assessment*, *23*(6), 837–849. https://doi.org/10.1007/s00477-008-0275-x

Kilsby, C. G., Jones, P. D., Burton, A., Ford, A. C., Fowler, H. J., Harpham, C., … Wilby, R. L. (2007). A daily weather generator for use in climate change studies. *Environmental Modelling and Software*, *22*(12), 1705–1719. https://doi.org/10.1016/j.envsoft.2007.02.005

Klemeš, V., & Borůvka, L. (1974). Simulation of Gamma-Distributed First-Order Markov Chain. *Water Resources Research*, *10*(1), 87–91. https://doi.org/10.1029/WR010i001p00087

Kolmogorov, A. N. (1940). Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen Raum. In *CR (Dokl.) Acad. Sci. URSS* (Vol. 26, pp. 115–118).

Kossieris, P., Makropoulos, C., Onof, C., & Koutsoyiannis, D. (2016). A rainfall disaggregation scheme for sub-hourly time scales: Coupling a Bartlett-Lewis based model with adjusting procedures. *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2016.07.015

Koutsoyiannis, D. (1999). Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology. *Water Resources Research*, *35*(4), 1219–1229.

https://doi.org/10.1029/1998WR900093

Koutsoyiannis, D. (2000). A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. *Water Resources Research*, *36*(6), 1519–1533. https://doi.org/10.1029/2000WR900044

Koutsoyiannis, D. (2002). The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal*, *47*(4), 573–595. https://doi.org/10.1080/02626660209492961

Koutsoyiannis, D. (2003). Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences Journal*, *48*(1), 3–24. https://doi.org/10.1623/hysj.48.1.3.43481

Koutsoyiannis, D. (2010). A random walk on water. *Hydrology and Earth System Sciences*, *14*(3), 585–601. https://doi.org/10.5194/hess-14-585-2010

Koutsoyiannis, D. (2011). Hurst-Kolmogorov Dynamics and Uncertainty. *JAWRA Journal of the American Water Resources Association*, *47*(3), 481–495. https://doi.org/10.1111/j.1752-1688.2011.00543.x

Koutsoyiannis, D. (2016). Generic and parsimonious stochastic modelling for hydrology and beyond. *Hydrological Sciences Journal*, *61*(2), 225–244. https://doi.org/10.1080/02626667.2015.1016950

Koutsoyiannis, D. (2017). Entropy Production in Stochastics. *Entropy*, *19*(11), 581. https://doi.org/10.3390/e19110581

Koutsoyiannis, D., Dimitriadis, P., Lombardo, F., & Stevens, S. (2018). From Fractals to Stochastics: Seeking Theoretical Consistency in Analysis of Geophysical Data. In *Advances in Nonlinear Geosciences* (pp. 237–278). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58895-7_14

Koutsoyiannis, D., & Economou, A. (2003). Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems. *Water Resources Research*, *39*(6), n/a-n/a. https://doi.org/10.1029/2003WR002148

Koutsoyiannis, D., & Foufoula-Georgiou, E. (1993). A scaling model of a storm hyetograph. *Water Resources Research*, *29*(7), 2345–2361. https://doi.org/10.1029/93WR00395

Koutsoyiannis, D., & Manetas, A. (1996). Simple disaggregation by accurate adjusting procedures. *Water Resources Research*, *32*(7), 2105–2117. https://doi.org/10.1029/96WR00488

Koutsoyiannis, D., & Montanari, A. (2007). Statistical analysis of hydroclimatic time series: Uncertainty and insights. *Water Resources Research*, *43*(5), 1–9. https://doi.org/10.1029/2006WR005592

Koutsoyiannis, D., & Papalexiou, S. M. (2016). Extreme rainfall: Global perspective. In *Chow's handbook of applied hydrology, 2nd Ed., McGraw-Hill, New York.*

Kroll, C. N., & Vogel, R. M. (2002). Probability Distribution of Low Streamflow Series in the United States. *Journal of Hydrologic Engineering*, *7*(2), 137–146. https://doi.org/10.1061/(ASCE)1084-0699(2002)7:2(137)

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical*

Association, *53*(284), 814–861.

Lawrance, A. J., & Lewis, P. A. W. (1981). A new autoregressive time series model in exponential variables (NEAR (1)). *Advances in Applied Probability*, *13*(04), 826–845.

Lebrun, R., & Dutfoy, A. (2009). An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, *24*(3), 312–320. https://doi.org/10.1016/j.probengmech.2008.08.001

Lee, T. (2016). Stochastic simulation of precipitation data for preserving key statistics in their original domain and application to climate change analysis. *Theoretical and Applied Climatology*, *124*(1–2), 91–102. https://doi.org/10.1007/s00704-015-1395-0

Lee, T. (2017). Multisite stochastic simulation of daily precipitation from copula modeling with a gamma marginal distribution. *Theoretical and Applied Climatology*. https://doi.org/10.1007/s00704-017-2147-0

Leonov, S., & Qaqish, B. (2017). Correlated endpoints: simulation, modeling, and extreme correlations. *Statistical Papers*. https://doi.org/10.1007/s00362-017-0960-2

Lettenmaier, D. P., & Burges, S. J. (1977). An operational approach to preserving skew in hydrologic models of long-term persistence. *Water Resources Research*, *13*(2), 281–290. https://doi.org/10.1029/WR013i002p00281

Li, S. T., & Hammond, J. L. (1975). Generation of Pseudorandom Numbers with Specified Univariate Distributions and Correlation Coefficients. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-5*(5), 557–561. https://doi.org/10.1109/TSMC.1975.5408380

Lindgren, G. (2013). *Stationary Stochastic Processes for Scientists and Engineers*. Chapman and Hall/CRC. https://doi.org/10.1201/b15922

Liu, P. L., & Der Kiureghian, A. (1986). Multivariate distribution models with prescribed marginals and covariances. *Probabilistic Engineering Mechanics*, *1*(2), 105–112. https://doi.org/10.1016/0266-8920(86)90033-0

Lombardo, F., Volpi, E., & Koutsoyiannis, D. (2012). Rainfall downscaling in time: theoretical and empirical comparison between multifractal and Hurst-Kolmogorov discrete random cascades. *Hydrological Sciences Journal*, *57*(6), 1052–1066. https://doi.org/10.1080/02626667.2012.695872

Lombardo, F., Volpi, E., Koutsoyiannis, D., & Papalexiou, S. M. (2014). Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology. *Hydrology and Earth System Sciences*, *18*(1), 243–255. https://doi.org/10.5194/hess-18-243-2014

Lombardo, F., Volpi, E., Koutsoyiannis, D., & Serinaldi, F. (2017). A theoretically consistent stochastic cascade for temporal disaggregation of intermittent rainfall. *Water Resources Research*. https://doi.org/10.1002/2017WR020529

Maftei, C., Barbulescu, A., & Carsteanu, A. A. (2016). Long-range dependence in the time series of Taiţa River discharges. *Hydrological Sciences Journal*, *61*(9), 1740–1747. https://doi.org/10.1080/02626667.2016.1171869

Mandelbrot, B. (1971). A Fast Fractional Gaussian Noise Generator. *Water Resources*

Research, *7*(3), 543–553. https://doi.org/10.1029/WR007i003p00543

Mandelbrot, B., & Wallis, J. R. (1969a). Computer Experiments With Fractional Gaussian Noises: Part 1, Averages and Variances. *Water Resources Research*, *5*(1), 228–241. https://doi.org/10.1029/WR005i001p00228

Mandelbrot, B., & Wallis, J. R. (1969b). Computer Experiments with Fractional Gaussian Noises: Part 2, Rescaled Ranges and Spectra. *Water Resources Research*, *5*(1), 242–259. https://doi.org/10.1029/WR005i001p00242

Mandelbrot, B., & Wallis, J. R. (1969c). Computer Experiments with Fractional Gaussian Noises: Part 3, Mathematical Appendix. *Water Resources Research*, *5*(1), 260–267. https://doi.org/10.1029/WR005i001p00260

Matalas, N. . C., & Wallis, J. R. (1976). *Generation of synthetic flow sequences, Systems Approach to Water Management*. (A. K. Biswas, Ed.). New York: McGraw-Hill, New York.

Matalas, N. C. (1967). Mathematical assessment of synthetic hydrology. *Water Resources Research*, *3*(4), 937–945. https://doi.org/10.1029/WR003i004p00937

Matalas, N. C., & Wallis, J. R. (1971). Statistical Properties of Multivariate Fractional Noise Processes. *Water Resources Research*, *7*(6), 1460–1468. https://doi.org/10.1029/WR007i006p01460

McMahon, T. A., Vogel, R. M., Peel, M. C., & Pegram, G. G. S. (2007). Global streamflows - Part 1: Characteristics of annual streamflows. *Journal of Hydrology*, *347*(3–4), 243–259. https://doi.org/10.1016/j.jhydrol.2007.09.002

Mehrotra, R., Li, J., Westra, S., & Sharma, A. (2015). A programming tool to generate multi-site daily rainfall using a two-stage semi parametric model. *Environmental Modelling and Software*, *63*, 230–239. https://doi.org/10.1016/j.envsoft.2014.10.016

Mejia, J. M., Rodriguez-Iturbe, I., & Dawdy, D. R. (1972). Streamflow simulation: 2. The broken line process as a potential model for hydrologic simulation. *Water Resources Research*, *8*(4), 931–941. https://doi.org/10.1029/WR008i004p00931

Mhanna, M., & Bauwens, W. (2012). A stochastic space-time model for the generation of daily rainfall in the Gaza Strip. *International Journal of Climatology*, *32*(7), 1098–1112. https://doi.org/10.1002/joc.2305

Molz, F. J., Liu, H. H., & Szulga, J. (1997). Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions. *Water Resources Research*, *33*(10), 2273–2286. https://doi.org/10.1029/97WR01982

Montanari, A., Rosso, R., & Taqqu, M. S. (1997). Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. *Water Resources Research*, *33*(5), 1035. https://doi.org/10.1029/97WR00043

Montanari, A., Rosso, R., & Taqqu, M. S. (2000). A seasonal fractional ARIMA Model applied to the Nile River monthly flows at Aswan. *Water Resources Research*, *36*(5), 1249–1259. https://doi.org/10.1029/2000WR900012

Mostafa, M. D., & Mahmoud, M. W. (1964). On the problem of estimation for the bivariate lognormal distribution. *Biometrika*, *51*(3–4), 522–527. https://doi.org/10.1093/biomet/51.3-4.522

Nataf, A. (1962). Statistique mathematique-determination des distributions de probabilites dont les marges sont donnees. *C. R. Acad. Sci. Paris*, *255*(1), 42–43.

Nazemi, A., Wheater, H. S., Chun, K. P., & Elshorbagy, A. (2013). A stochastic reconstruction framework for analysis of water resource system vulnerability to climate-induced changes in river flow regime. *Water Resources Research*, *49*(1), 291–305. https://doi.org/10.1029/2012WR012755

O'Connell, P. E., Koutsoyiannis, D., Lins, H. F., Markonis, Y., Montanari, A., & Cohn, T. (2016). The scientific legacy of Harold Edwin Hurst (1880–1978). *Hydrological Sciences Journal*, *61*(9), 1571–1590. https://doi.org/10.1080/02626667.2015.1125998

Papalexiou, S. M. (2018). Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*. https://doi.org/10.1016/j.advwatres.2018.02.013

Papalexiou, S. M., & Koutsoyiannis, D. (2013). Battle of extreme value distributions : A global survey on extreme daily rainfall. *Water Resources Research*. https://doi.org/10.1029/2012WR012557

Papalexiou, S. M., & Koutsoyiannis, D. (2016). A global survey on the seasonal variation of the marginal distribution of daily precipitation. *Advances in Water Resources*, *94*, 131–145. https://doi.org/10.1016/j.advwatres.2016.05.005

Papalexiou, S. M., Koutsoyiannis, D., & Makropoulos, C. (2013). How extreme is extreme? An assessment of daily rainfall distribution tails. *Hydrology and Earth System Sciences*, *17*(2), 851–862. https://doi.org/10.5194/hess-17-851-2013

Papalexiou, S. M., Koutsoyiannis, D., & Montanari, A. (2011). Can a simple stochastic model generate rich patterns of rainfall events? *Journal of Hydrology*, *411*(3–4), 279–289. https://doi.org/10.1016/j.jhydrol.2011.10.008

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes* (Third edit). McGraw-Hill Series in Electrical Engineering. New York City, New York, USA: McGraw-Hill.

Paschalis, A., Fatichi, S., Molnar, P., Rimkus, S., & Burlando, P. (2014). On the effects of small scale space–time variability of rainfall on basin flood response. *Journal of Hydrology*, *514*, 313–327. https://doi.org/10.1016/j.jhydrol.2014.04.014

Pegram, G. G. S., & James, W. (1972). Multilag multivariate autoregressive model for the generation of operational hydrology. *Water Resources Research*, *8*(4), 1074–1076. https://doi.org/10.1029/WR008i004p01074

Pui, A., Sharma, A., Mehrotra, R., Sivakumar, B., & Jeremiah, E. (2012). A comparison of alternatives for daily to sub-daily rainfall disaggregation. *Journal of Hydrology*, *470–471*, 138–157. https://doi.org/10.1016/j.jhydrol.2012.08.041

Qin, X. S., & Lu, Y. (2014). Study of Climate Change Impact on Flood Frequencies: A Combined Weather Generator and Hydrological Modeling Approach. *Journal of Hydrometeorology*, *15*(3), 1205–1219. https://doi.org/10.1175/JHM-D-13-0126.1

Salas, J. D., Delleur, J. W., Yevjevich, V., & Lane, W. L. (1980). *Applied modeling of hydrologic time series* (2nd Print). Littleton, Colorado: Water Resources Publication.

Samoradnitsky, G. (2017). *Stable non-Gaussian random processes: stochastic models with infinite variance*. Routledge.

Serinaldi, F. (2009). A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. *Journal of Geophysical Research*, *114*(D10), D10103. https://doi.org/10.1029/2008JD011258

Serinaldi, F., & Kilsby, C. G. (2014). Simulating daily rainfall fields over large areas for collective risk estimation. *Journal of Hydrology*, *512*, 285–302. https://doi.org/10.1016/j.jhydrol.2014.02.043

Serinaldi, F., & Lombardo, F. (2017). BetaBit: A fast generator of autocorrelated binary processes for geophysical research. *EPL (Europhysics Letters)*, *118*(3), 30007. https://doi.org/10.1209/0295-5075/118/30007

Srikanthan, R., & McMahon, T. a. (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, *5*(4), 653–670. https://doi.org/10.5194/hess-5-653-2001

Srikanthan, R., & Pegram, G. G. S. (2009). A nested multisite daily rainfall stochastic generation model. *Journal of Hydrology*, *371*(1–4), 142–153. https://doi.org/10.1016/j.jhydrol.2009.03.025

Stacy, E. W. (1962). A Generalization of the Gamma Distribution. *The Annals of Mathematical Statistics*, *33*(3), 1187–1192. https://doi.org/10.1214/aoms/1177704481

Tessier, Y., Lovejoy, S., Hubert, P., Schertzer, D., & Pecknold, S. (1996). Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. *Journal of Geophysical Research: Atmospheres*, *101*(D21), 26427–26440. https://doi.org/10.1029/96JD01799

Thomas, H. A., & Fiering, M. B. (1963). The nature of the storage yield function. *Operations Research in Water Quality Management*.

Thompson, C. S., Thomson, P. J., & Zheng, X. (2007). Fitting a multisite daily rainfall model to New Zealand data. *Journal of Hydrology*, *340*(1–2), 25–39. https://doi.org/10.1016/j.jhydrol.2007.03.020

Todini, E. (1980). The preservation of skewness in linear disaggregation schemes. *Journal of Hydrology*, *47*(3–4), 199–214. https://doi.org/10.1016/0022-1694(80)90093-1

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

Tsoukalas, I., Efstratiadis, A., & Makropoulos, C. (2018a). Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions. *Water Resources Research*, *54*(1), 161–185. https://doi.org/10.1002/2017WR021394

Tsoukalas, I., & Makropoulos, C. (2015a). A Surrogate Based Optimization Approach for the Development of Uncertainty-Aware Reservoir Operational Rules: the Case of Nestos

Hydrosystem. *Water Resources Management*, *29*(13), 4719–4734. https://doi.org/10.1007/s11269-015-1086-8

Tsoukalas, I., & Makropoulos, C. (2015b). Multiobjective optimisation on a budget: Exploring surrogate modelling for robust multi-reservoir rules generation under hydrological uncertainty. *Environmental Modelling & Software*, *69*, 396–413. https://doi.org/10.1016/j.envsoft.2014.09.023

Tsoukalas, I., Papalexiou, S., Efstratiadis, A., & Makropoulos, C. (2018b). A Cautionary Note on the Reproduction of Dependencies through Linear Stochastic Models with Non-Gaussian White Noise. *Water*, *10*(6), 771. https://doi.org/10.3390/w10060771

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference* (Vol. 579, p. 6o4).

Tyralis, H., & Koutsoyiannis, D. (2011). Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process. *Stochastic Environmental Research and Risk Assessment*, *25*(1), 21–33. https://doi.org/10.1007/s00477-010-0408-x

Wheater, H. S., Chandler, R. E., Onof, C. J., Isham, V. S., Bellone, E., Yang, C., … Segond, M.-L. (2005). Spatial-temporal rainfall modelling for flood risk estimation. *Stochastic Environmental Research and Risk Assessment*, *19*(6), 403–416. https://doi.org/10.1007/s00477-005-0011-8

Whitt, W. (1976). Bivariate Distributions with Given Marginals. *The Annals of Statistics*, *4*(6), 1280–1289. https://doi.org/10.1214/aos/1176343660

Wilks, D. S. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, *210*(1–4), 178–191. https://doi.org/10.1016/S0022-1694(98)00186-3

Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, *23*(3), 329–357. https://doi.org/10.1191/030913399666525256

Xiao, Q. (2014). Evaluating correlation coefficient for Nataf transformation. *Probabilistic Engineering Mechanics*, *37*, 1–6. https://doi.org/10.1016/j.probengmech.2014.03.010