# Stochastic periodic autoregressive to anything (SPARTA): Modelling and simulation of cyclostationary processes with arbitrary marginal distributions

**Ioannis Tsoukalas**[1,*], **Andreas Efstratiadis**[1], **and Christos Makropoulos**[1]

[1] 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 periodic processes with any marginal distributions
- Flexibility in the selection of distribution fitting method
- Generation of synthetic time series in univariate or multivariate mode
- Accurate preservation of essential statistics and observed dependencies

**Keywords:**

stochastic simulation, hydrological processes, cyclostationarity, synthetic time series, multivariate autoregressive models, arbitrary marginal distributions, Nataf joint-distribution model, normal to anything, linear correlation, dependence patterns

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

**Supporting information:**

http://onlinelibrary.wiley.com/store/10.1002/2017WR021394/asset/supinfo/wrcr23047-sup-0001-FigS01.pdf?v=1&s=09a47c3b688bf7e9ef4e01f72aacca90c6cc5f32

**Citation:**

Tsoukalas, I., Efstratiadis, A., & Makropoulos, C. (2017). Stochastic periodic autoregressive to anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions. Water Resources Research, 53. https://doi.org/10.1002/2017WR021394

**Journal:** Water Resources Research

Received 26 JUN 2017

Accepted 17 DEC 2017

Accepted article online 21 DEC 2017

## Abstract

Stochastic models in hydrology traditionally aim at reproducing the empirically-derived statistical characteristics of the observed data rather than any specific distribution model that attempts to describe the usually non-Gaussian statistical behavior of the associated processes. SPARTA (Stochastic Periodic AutoRegressive To Anything) offers an alternative and novel approach which allows the explicit representation of each process of interest with any distribution model, while simultaneously establishes dependence patterns that cannot be fully captured by the typical linear stochastic schemes. Cornerstone of the proposed approach is the Nataf joint-distribution model, which is related with the Gaussian copula, combined with Gaussian periodic autoregressive processes. In order to obtain the target stochastic structure, we have also developed a computationally simple and efficient algorithm, based on a hybrid Monte-Carlo procedure that is used to approximate the required equivalent correlation coefficients. Theoretical and practical benefits of the proposed method, contrasted to outcomes from widely-used stochastic models, are demonstrated by means of real-world as well as hypothetical monthly simulation examples involving both univariate and multivariate time series.

## Plain Language Summary

Stochastic hydrology and, particularly, the synthesis of long hydrometeorological time series (e.g., precipitation and streamflow) is of high importance in water-related studies since it enables to account for the intrinsic uncertainty of the associated processes. This in turn provides the means to embed uncertainty within planning and decision making. Typically, stochastic models in hydrology aim in the resemblance of the empirically-derived statistical characteristics of the observed time series rather than in reproducing a specific distribution model. In this work we propose a novel approach termed SPARTA (Stochastic Periodic AutoRegressive To Anything) that allows the simulation of multivariate cyclostationary processes with explicit reproduction of the desirable marginal distributions and correlation structures. Its theoretical background is based on the Nataf joint-distribution model (NDM), a procedure that emerged from operations research and is also related with the Gaussian copula. The theoretical and practical benefits of the proposed method are demonstrated by means of real-world and hypothetical simulation studies, involving the generation of both univariate and multivariate time series.

## 1 Introduction

According to the classical classification by *Matalas* [1975], synthetic hydrology constitutes a sub-branch of stochastic hydrology, which is usually credited to the pivotal works conducted by the Harvard water program [*Maass et al.,* 1962] and *Thomas and Fiering* [1962]. Early attempts to simulate synthetic time series were based on the theory of stochastic processes and the use of linear stochastic models, accounting for the key peculiarities of hydrometeorological processes, namely periodicity and skewness [*Thomas and Burden,* 1963; *Matalas,* 1967; *Fiering and Jackson,* 1971; *Klemeš and Borůvka,* 1974].

Typically, the standard hypothesis for synthetic time series generation via such approaches does not lie in the reproduction of a specific distribution, but on the resemblance of the statistical characteristics of the parent historical time series. These are usually expressed in terms of low-order statistics (e.g. mean, variance, skewness) and correlations in time and space [*Matalas and Wallis,* 1976; *Salas,* 1993]. However, for a given set of low-order statistics multiple distribution functions may be represented, thus making the simulation problem only partially defined [cf. *Matalas and Wallis*, 1976 p. 66].

On the other hand, theoretical reasons and empirical evidence may impose the preservation of a specific distribution for the modelled processes, as highlighted by *Klemeš and Borůvka* [*1974*] (our emphasis):

> "*Simulation of a serially correlated series* ***with a given marginal distribution*** *is one of the important prerequisites of synthetic hydrology and of its applications to analysis of water resource system*".

The generation of synthetic data following specific, typically skewed, distributions becomes even more challenging when aiming to simulate hydrometeorological processes at time scales finer than annual. In that case, the stochastic model should account for all facets of cyclostationarity, involving not only, the stochastic structure of the underlying processes but also their distribution, which may be seasonally-varying.

The standard approaches to handle skewness within linear stochastic models can be classified in three categories: (a) explicit methods, (b) transformation methods, and c) implicit methods that produce non-Gaussian innovation terms within the generation scheme. Such approaches suffer from notable, although not so apparent, limitations that in combination with the need to account for non-Gaussian distributions motivated this study.

Explicit methods are designed (and hence constrained) to generate realizations from a specific distribution family. Common approaches within stochastic hydrology are the stationary multivariate lag-1 model with Log-Normal distribution, proposed by *Matalas* [1967], and the gamma-autoregressive (GAR) model of *Lawrance and Lewis* [1981], as well as its periodic extension [*Fernandez and Salas, 1986*]. We remark that so far GAR is restricted for univariate cases, which is a major limitation, since in most water resources applications multiple processes have to be represented simultaneously.

Transformation approaches initially aim to "normalize" the non-Gaussian historical data through a proper transformation function; next, parameter estimation and simulation are performed on the normalized data and the final product, i.e., the synthetic data, are obtained via the inverse transformation [*Salas et al., 1985*]. Early attempts used relatively simple conversions, such as Box-Cox, logarithmic, and alternatives, which is well-known that cannot always ensure a satisfactory normalization (e.g. when the original data are too asymmetric). For this reason, for the case of hydrometeorological data, exhibiting significant skewness, more complex schemes have been proposed, yet involving several unknown parameters and also requiring the use of optimization [e.g., *Koutsoyiannis et al.*, 2008; *Papalexiou et al.*, 2011]. In fact, the increase of complexity inevitably raises several questions, namely: How many parameters should be used? How does the sample size affect their estimation? In the case of multivariate and cyclostationary simulations, should we use the same transformation function for all processes and seasons?

Nevertheless, even an accurate normalization procedure does not ensure that the inverse transformation (i.e., the normalization – simulation – de-normalization scheme) will preserve both the statistical characteristics and the correlation structure of the original variables [*Salas et al.*, 1980 p. 73; *Bras and Rodríguez-Iturbe*, 1985; *Lall and Sharma*, 1996; *Sharma et al.*, 1997]. Actually, it is argued that a general method for normalizing all types of data does not exist [*Papalexiou et al.*, 2011]. We could also argue that neither an optimal transformation for each specific process exists (particularly in the multivariate case). Thus, the selection and the parameters of the transformation model are prone to subjectivity and indefiniteness.

To avoid such ill-transformations, the common practice has leaned towards incorporating skewness within the generation mechanism of the stochastic model itself. In this context, several *implicit* schemes have been proposed to embed non-Gaussian noise within the innovation term. The first attempts are attributed to *Thomas and Burden* [1963] and *Fiering and Jackson* [1971] who proposed univariate simulation schemes for skewed and periodic streamflow data. Their key assumption is the preservation of the desirable statistical characteristics through the generation of white noise from a given distribution, usually the three-parametric Gamma (Pearson type-III). We remark that such approaches generate *explicitly* gamma-distributed variables for the white noise, while the strict "explicitness" is lost when the latter are synthesized to provide the variables of interest [cf. *Matalas and Wallis,* 1976 p. 66]. Hence, the desirable distribution is only approximately preserved [*Koutsoyiannis and Manetas,* 1996]. After the pioneering works of *Fiering* [1964] and *Matalas and Wallis* [1976], implicit approaches have been implemented within several linear stochastic models, including the multivariate periodic autoregressive model [cf. *Koutsoyiannis,* 1999], the multivariate symmetric moving average model [*Koutsoyiannis,* 2000] and their integration within Castalia framework [*Efstratiadis et al.*, 2014].

A well-known alternative to all above categories of linear stochastic models is offered by the so-called *non-parametric* approaches, which aim to reproduce the empirical distributions of the observed processes, typically through resampling of historical data [e.g., *Lall and Sharma*, 1996; *Sharma et al.*, 1997; *Srinivas and Srinivasan,* 2005; *Mehrotra et al.*, 2006; *Marković et al.*, 2015]. In the literature, such approaches have gained particular attention when the marginal distributions exhibit bi- or multi-modality, which is usually driven by multiple generation mechanisms [*Lall and Sharma*, 1996; *Sharma et al.*, 1997]. However, the use of the empirical distributions prohibits from fitting to a theoretical model and extrapolating out of the observed data ranges. The lack of theoretical basis makes also difficult to reproduce long-term persistence and cross-correlations among many variables, with few exceptions [e.g., *Kirsch et al., 2013; Herman et al.*, 2016]. Heuristic solutions to the above limitations, such as the recently introduced optimization-based approach by *Borgomeo et al.* [2015], are subject to extremely high computational effort, and they are also prone to inherent inefficiencies of optimization algorithms.

Another relatively new and promising option is offered by copulas, which have recently been embedded in multivariate stochastic simulation schemes in order to describe complex dependencies among hydrological variables [*Hao and Singh*, 2013; *Chen et al.*, 2015]. However, it can be argued that copulas are not directly compatible with linear stochastic models, which rely on Pearson's correlation coefficient, since they typically employ rank correlation statistics (e.g., Spearman's $\rho_s$, or Kendall's $\tau$) to describe the dependencies among the variables. Nevertheless, they are more sensitive against sampling uncertainty than classical stochastic schemes, in their attempt to describe complex (i.e., nonlinear) dependencies on the basis of usually limited hydrological data. Furthermore, as many researchers argue (see discussion in the aforementioned papers), they rely on quite complicated and computationally demanding generation schemes, especially in high-dimensional spaces.

In order to tackle the aforementioned shortcomings, we have developed an *explicit* method, called Stochastic Periodic AutoRegressive To Anything (SPARTA) model, which offers a *generalized* procedure with solid theoretical background for the generation of cyclostationary processes from *a priori* defined distribution functions that are seasonally-varying. The proposed method builds upon the so-called Nataf joint-distribution model [NDM; *Nataf*, 1962], which is generic mapping procedure, and the AutoRegressive To Anything (ARTA) model, introduced by *Cario and Nelson* [1996] to represent stationary processes with arbitrary marginal distributions and autocorrelation structure. Initially, ARTA was formulated as univariate and later extended for multivariate simulations [*Biller and Nelson*, 2003]. Both versions involve the simulation of stationary processes, but they have not been extended to account for cyclostationarity which is *sine qua non* requirement for hydrological processes.

The rationale and computational procedure of SPARTA are described in the next three sections, where section 2 summarizes the overall methodology, section 3 describes the individual computational steps, while section 4 emphasizes on the Nataf joint-distribution model and the associated numerical scheme that has been developed and implemented within SPARTA. In section 5 we evaluate the proposed method by means of three case studies, involving real-world and hypothetical simulations. A broader discussion on good modelling practices, as well as the key conclusions and perspectives of this research are outlined in sections 6 and 7, respectively.

## 2 SPARTA at a glance

SPARTA aims at simulating periodic processes from any given marginal distribution and a given stochastic structure, typically (but not exclusively) expressed in terms of first order autocorrelations and lag zero cross-correlations. Its fundamental advantage is the explicit preservation of the theoretical marginal distributions of the processes, in contrast to existing linear stochastic approaches that preserve the marginal statistics (not the distributions themselves) up to a specific order, typically the third one (i.e., mean, standard deviation, skewness). Briefly, our approach involves the simulation of an auxiliary process from the Periodic AutoRegressive (PAR) family, in the "normal" domain (i.e., Gaussian), which allows accounting for cyclostationarity, and then its mapping to the "real" domain, via the desired inverse cumulative distribution functions (ICDFs). More specifically: Let $\underline{\boldsymbol{x}}_{s,t} = \left[\underline{x}_{s,t}^{1}, \dots, \underline{x}_{s,t}^{m}\right]^{\mathrm{T}}$ be a $m$-dimensional vector of cyclostationary stochastic processes to simulate, where $s = 1, \dots, S$ denotes the season (e.g., month) and $t = 1, \dots, T$ denotes the aggregated time scale (e.g., year). Each element of $\underline{\boldsymbol{x}}_{s,t}$ is symbolized $\underline{x}_{s,t}^{i}$, where $i = 1, \dots, m$ denotes an individual random process, and $x_{s,t}^{i}$ denotes its realization. Herein, index $i$ will be also referred to as "location" or "site", without necessarily implying spatial reference. Let also $\rho_{s,s-\tau}^{i,j} \coloneqq \mathrm{Corr}[\underline{x}_{s}^{i}, \underline{x}_{s-\tau}^{j}]$ be the Pearson coefficient of correlation among processes $i$ and $j$, for season $s$ and time lag $\tau$. For instance, when $j = i$ and $\tau \neq 0$, the quantity $\rho$ represents the autocorrelation of the process for lag $\tau$, while for $j \neq i$ and $\tau = 0$, $\rho$ represents the cross-correlation between $i$ and $j$, for zero time lag. Furthermore, when the superscripts or subscripts of $\rho$ are identical (i.e., when $j = i$ or $\tau = 0$) we may omit repeating them for convenience (e.g., $\rho_{s,s-\tau}^{i,i}$ may be written as $\rho_{s,s-\tau}^{i}$ and $\rho_{s,s}^{i,j}$ as $\rho_{s}^{i,j}$).

For each process at each season $s$ and each location $i$, we assign a specific statistical distribution, $F_{\underline{x}_{s}^{i}} \coloneqq P(\underline{x}_{s}^{i} \leq \mathrm{x})$, and also assign target coefficients of correlation, $\rho_{s,s-\tau}^{i,j}$, to preserve within the proposed generation scheme. The key idea of SPARTA lies in the generation of an auxiliary process $\underline{\boldsymbol{z}}_{s,t} = \left[\underline{z}_{s,t}^{1}, \dots, \underline{z}_{s,t}^{m}\right]^{\mathrm{T}}$ from a standard Normal Periodic AutoRegressive process (symbolized PAR-N), with such parameters that their mapping via the corresponding inverse marginal distributions (ICDFs) results into processes with the target marginal distributions and the target correlation structure, i.e.,

$$\underline{x}_{s,t}^{i} = F_{\underline{x}_s^i}^{-1}[\Phi(\underline{z}_{s,t}^{i})] \tag{1}$$

where $\Phi(\cdot)$ is the CDF of the standard Gaussian distribution and $F_{\underline{x}_s^i}^{-1}(\cdot)$ denotes the ICDFs of the target distributions of process *i* at season *s*.

The main challenge, also encountered in the original model (i.e., ARTA), is the identification of proper parameters for the auxiliary process in the "normal" domain that reproduce the desired stochastic structure, after applying the mapping procedure. This arises from the fact that the Pearson correlation coefficient, which is used to describe all kinds of dependencies within linear stochastic models (including PAR), cannot be preserved when applying a non-linear monotonic transformation, such as the ICDF. In particular, Eq. (1) results into underestimation of target correlations, $\rho_{s,s-\tau}^{i,j}$, when they are directly applied to the auxiliary processes. The origin of this shortcoming is the fact that the Pearsons' correlation coefficient (in contrast to rank correlation statistics) is invariant only under linear transformations [*Embrechts et al.*, 1999 p. 7], while for any other transformation, the correlation coefficients should be properly adjusted. As we will discuss later (section 4.1), early works in stochastic hydrology were aware of this issue and attempted to provide analytical or empirical solutions to this problem, for specific distributions (e.g., Log-Normal).

Following the rationale of ARTA, here we ensure the representation of *any* distribution across seasons and processes by employing the so-called Nataf joint-distribution model [NDM; *Nataf*, 1962]. NDM offers a generic solution to the mapping problem, thus assigning suitable coefficients to the auxiliary processes that will finally attain the desirable correlation after the transformation to the "real" domain. Here, we employ NDM in order to identify such "equivalent" coefficients, $\tilde{\rho}_{s,s-\tau}^{i,j}$, to be used within the PAR-N generation procedure. As will be elucidated in section 4, for their estimation we have developed a hybrid method, on the basis of target CDFs, $F_{\underline{x}_s^i}$, and target $\rho_{s,s-\tau}^{i,j}$.

Summarizing, the implementation of SPARTA comprises five steps:

**Step 1**: For each variable *i* and each season *s*, specify a suitable target marginal distribution, $F_{\underline{x}_s^i}$, and also identify the dependencies to be preserved in time and space, as well as the target values of the associated coefficients of correlation, $\rho_{s,s-\tau}^{i,j}$.

**Step 2**: On the basis of the desirable dependencies to preserve (in terms of auto- and cross-correlations), identify the suitable auxiliary model from the PAR-N family.

**Step 3**: Employ NDM to determine the equivalent coefficients of correlation, $\tilde{\rho}_{s,s-\tau}^{i,j}$, for all pairs of variables that are required by the auxiliary model.

**Step 4**: Estimate the parameters of the auxiliary model, on the basis of equivalent correlations, and run the model to generate the auxiliary Gaussian synthetic time series of $\underline{\boldsymbol{z}}_{s,t}$.

**Step 5**: Map the auxiliary process $\underline{\boldsymbol{z}}_{s,t}$ to the actual domain using their ICDFs, i.e., through Eq. (1), to obtain $\underline{\boldsymbol{x}}_{s,t}$.

The above steps are described in section 3, while step 3, which is the core element of the proposed methodology, is discussed in detail in section 4.

# 3 Insights to the computational procedure

## 3.1 Selection of target marginal distributions and correlations

In contrast to classical stochastic approaches, which imply the use of a specific statistical model for the noise, SPARTA allows to employ pre-specified distribution models, in order to describe the statistical structure of the modelled processes themselves and not of the noise, which is an auxiliary process. This flexibility involves the selection of the marginal distributions, $F_{\underline{x}_s^i}$, and the identification of their parameters. In addition, the proposed approach allows for identifying target dependencies to preserve, in time and space, expressed by means of target coefficients of correlation, $\rho_{s,s-\tau}^{i,j}$. We highlight that the specification of the above inputs is not a straightforward decision neither it is advised to be made automatically. As thoroughly discussed in section 6, the modeler should account for multilateral information, based both on historical data and expert judgment, in order to establish a realistic formulation of the stochastic simulation model.

## 3.2 The auxiliary model

As mentioned above, the generation procedure of SPARTA requires the synthesis of an auxiliary process $\underline{\boldsymbol{z}}_{s,t}$, which is then mapped to the actual one, i.e., $\underline{\boldsymbol{x}}_{s,t}$. This process has to be cyclostationary (since the underlying process is also cyclostationary) and normal. These premises are fulfilled by standard periodic autoregressive models with normally-distributed noise (PAR-N) of any order [e.g., *Salas and Pegram*, 1977; *Salas et al.*, 1985; *Salas*, 1993].

Although any stochastic scheme from the PAR-N family may be applicable, we pay attention to the PAR(1) process, in order to keep things simple and parsimonious, thus providing an easy to follow narrative. In addition, it is argued that the assumption of a first-order model is well-justified for most of practical applications in hydrology [*Efstratiadis et al.*, 2014]. Nevertheless, higher-order models may be cumbersome, because the empirical estimation of joint statistics from historical samples is subject to major uncertainty, usually resulting to ill-posed conditions (e.g., due to inconsistent autocorrelation structures), which in turn leads to substantial defects within parameter estimation.

With respect to cross-correlations, the multivariate PAR(1) model, in its full formulation, preserves both the lag zero and lag one dependencies. However, as *Koutsoyiannis and Manetas* [1996] have shown, for reasons of parsimony it is sufficient using the contemporaneous PAR(1) [*Salas*, 1993 p. 19.31], which does not explicitly accounts for lag-one cross-correlations within parameter estimation. This is also advocated by an older study of *Pegram and James* [1972]. For instance, in a four-variable problem with 12 seasons, the full PAR(1) model requires the specification of 264 parameters to describe the dependencies among the variables, while the contemporaneous one entails 120.

## 3.3 Estimation of equivalent coefficients of correlation

In order to employ the multivariate contemporaneous PAR(1)-N within SPARTA, it is essential to provide the equivalent lag-1 month-to-month correlations (i.e., autocorrelations), $\tilde{\rho}_{s,s-1}^{i}$, for each process *i* and season *s*, as well as the equivalent zero-lag cross-correlations, $\tilde{\rho}_{s}^{i,j}$, for each pair of processes *i* and *j* and season *s*. We remark that the equivalent correlations differ from the target ones, and they are estimated on the basis of the NDM approach, which is described in detail in section 4.

### 3.4 Parameter estimation within PAR(1)-N process

#### 3.4.1 Multivariate contemporaneous case

Keeping the same notation for the auxiliary and actual processes, the multivariate PAR(1) reads (for convenience, time index $t$ is omitted):

$$\underline{\boldsymbol{z}}_s = \widetilde{\boldsymbol{A}}_s \underline{\boldsymbol{z}}_{s-1} + \widetilde{\boldsymbol{B}}_s \underline{\boldsymbol{w}}_s \tag{2}$$

where $\underline{\boldsymbol{z}}_s = \left[\underline{z}_s^1, \dots, \underline{z}_s^m\right]^{\mathrm{T}}$ is a vector of $m$ stochastic processes in season $s$, $\widetilde{\boldsymbol{A}}_s, \widetilde{\boldsymbol{B}}_s$ are $m \times m$ parameter matrices that depend on season $s$, and $\underline{\boldsymbol{w}}_s = \left[\underline{w}_s^1, \dots, \underline{w}_s^m\right]^{\mathrm{T}}$ is a vector of mutually independent random variables. By definition, the random process $\underline{\boldsymbol{z}}_s$ is Gaussian, provided that $\underline{\boldsymbol{w}}_s$ is generated from the standard normal distribution, i.e., $\underline{\boldsymbol{w}}_s \sim N(0, 1)$.

For each season $s$, the parameter matrix $\widetilde{\boldsymbol{A}}_s$ is diagonal and contains the equivalent lag-1 month-to-month correlations, $\tilde{\rho}_{s,s-1}^i$, i.e.,

$$\widetilde{\boldsymbol{A}}_s = \mathrm{diag}\left(\tilde{\rho}_{s,s-1}^1, \dots, \tilde{\rho}_{s,s-1}^m\right) \tag{3}$$

On the other hand, parameter matrices $\widetilde{\boldsymbol{B}}_s$ are calculated as follows:

$$\widetilde{\boldsymbol{B}}_s \widetilde{\boldsymbol{B}}^{\mathrm{T}}{}_s = \widetilde{\boldsymbol{G}}_s \tag{4}$$

where $\widetilde{\boldsymbol{G}}_s := \widetilde{\boldsymbol{C}}_s - \widetilde{\boldsymbol{A}}_s \widetilde{\boldsymbol{C}}_{s-1} \widetilde{\boldsymbol{A}}^{\mathrm{T}}{}_s$ and $\widetilde{\boldsymbol{C}}_s$ is a symmetric $m \times m$matrix that contains the equivalent lag-zero cross-correlations, $\tilde{\rho}_s^{i,j}$, i.e.,

$$\widetilde{\boldsymbol{C}}_s = \begin{pmatrix} 1 & \cdots & \tilde{\rho}_s^{1,m} \\ \vdots & \ddots & \vdots \\ \tilde{\rho}_s^{m,1} & \cdots & 1 \end{pmatrix}$$

In order to estimate the parameter matrix $\widetilde{\boldsymbol{B}}_s$, it is essential to solve a decomposition problem, also expressed as finding the square root of $\widetilde{\boldsymbol{G}}_s$. This can be obtained with the use of typical numerical techniques, such as Cholesky or singular value decomposition [e.g., *Johnson*, 1987]. We remark that when $\widetilde{\boldsymbol{G}}_s$ is positive definite, it has infinite number of feasible solutions, such as the solutions provided by the aforementioned numerical methods. On the other hand, if $\widetilde{\boldsymbol{G}}_s$ is non-positive definite (this is often the case when the historical data are of different length) the problem does not have a feasible solution, thus requiring the detection of a parameter matrix $\widetilde{\boldsymbol{B}}_s$ ensuring an approximation of the given $\widetilde{\boldsymbol{G}}_s$, e.g., through optimization [*Koutsoyiannis*, 1999; *Higham*, 2002].

In particular, *Koutsoyiannis* [1999] has developed an optimization-based approach, paying attention on the preservation of skewness, which is a well-known trouble of multivariate stochastic models, asking for generating skewed white noise [e.g., *Todini*, 1980]. A great advantage of our approach is the assumption of normality within the auxiliary process, which substantially simplifies the optimization problem within decomposing non-positive definite matrices. More precisely, the empirical penalty term considered by *Koutsoyiannis* [1999], in order to prohibit the generation of highly-skewed white noise, which introduces significant complexity to the optimization procedure [cf. *Efstratiadis et al.*, 2014], is neglected, thus resulting to a "reduced" objective function that only contains a distance term to minimize.

### 3.4.2 Univariate case

The univariate model can easily be derived from the above equations. Since $m$ = 1, $\widetilde{\boldsymbol{A}}_s = \tilde{\rho}^1_{s,s-1}$ and $\widetilde{\boldsymbol{C}}_s = 1$, thus $\widetilde{\boldsymbol{B}}_s \widetilde{\boldsymbol{B}}^{\mathrm{T}}{}_s = 1 - \tilde{\rho}^1_{s,s-1}\tilde{\rho}^1_{s,s-1}$, which leads to $\widetilde{\boldsymbol{B}}_s = \sqrt{1 - \tilde{\rho}^1_{s,s-1}{}^2}$. Hence, by substituting in Eq. (2) and removing the redundant indices we read:

$$\underline{z}_s = \tilde{\rho}_{s,s-1}\underline{z}_{s-1} + \sqrt{1 - \tilde{\rho}_{s,s-1}{}^2}\,\underline{w}_s \tag{5}$$

where $\underline{w}_s$ are i.i.d. white noise with $N{\sim}(0,1)$. We remark that since $i$ = 1 the superscript of $\tilde{\rho}(\cdot)$ has been omitted for simplicity.

## 3.5 Mapping auxiliary processes to the actual domain

After generating the synthetic time series of the auxiliary processes $\underline{\boldsymbol{z}}_s$, the last step is its mapping throughout Eq. (1) to the actual domain $\underline{\boldsymbol{x}}_s$, through the inverse CDFs. This procedure is implemented for each individual process and season. Due to the use of the inverse CDF, as well as the use of equivalent coefficients of correlation within the PAR(1)-N model, the resulting data will preserve both the target marginal distributions, for all seasons and locations, as well as the target auto- and cross-correlations. Even in case of non-positive definite correlation matrices, where the desired stochastic characteristics are not explicitly preserved by the PAR(1)-N model, the "reduced" optimization approach ensures a very good approximation, with minimal computational burden.

# 4 Nataf joint-distribution model and computational advances

## 4.1 Historical summary and rationale

The problem of obtaining a joint pdf of random variables based on their individual distributions and correlation has long been discussed within the statistical community. *Nataf* [1962] has proposed a quite simple, yet general solution by mapping multivariate normal variables with a given correlation matrix to multivariate uniform variables, which in turn are mapped to the desired distributions via the corresponding inverse cumulative functions. The key challenge is to identify the equivalent correlations to be applied within the generation of random variables in the normal domain, in order to attain the desired correlation in the real domain. In their classical work, *Liu and Der Kiureghian* [1986] showed that the Nataf's Distribution Model (NDM) is suitable for describing a wide range of correlation values. Later, *Cario and Nelson* [1997], developed a generalized procedure based on NDM and referred to as NORTA (NORmal To Anything), for the generation of correlated random vectors with arbitrary marginal distributions, including discrete and mixed ones. In fact, NDM may be considered as a specific case of copulas [*Sklar*, 1973], and more specifically the Gaussian one. In fact, linear stochastics are compatible with the latter copula, since both use the Pearson's linear correlation as measure of dependence. *Lebrun and Dutfoy* [2009], in view of copula theory, provide an extensive and insightful discussion on the relation of NDM with the Gaussian copula, as well as provide an alternative formulation of the former in terms of Spearman's $\rho_{\mathrm{s}}$ and Kendall's $\tau$.

We remark that when *Cario and Nelson* [1997] have published their work, they argued that the generality of their approach came at the cost of computational efficiency (i.e., computational time), since the estimation $\tilde{\rho}$ presupposed solving numerically a double integral in the infinite domain. However, this argument is far from interest now, grace to continuous advances in computing, which have significantly contributed in waiving such barriers.

### 4.2 Theoretical background

In the general case, let that we wish to generate a correlated random vector $\boldsymbol{\underline{x}} = [\underline{x}_1, \dots, \underline{x}_k, \dots, \underline{x}_m]^{\mathrm{T}}$ with target marginal distributions $F_{\underline{x}_k}$ and target correlation matrix:

$$\boldsymbol{C}_{\underline{\boldsymbol{x}}} = \begin{pmatrix} 1 & \cdots & \rho_{1,m} \\ \vdots & \ddots & \vdots \\ \rho_{m,1} & \cdots & 1 \end{pmatrix}$$

Let also $\boldsymbol{\underline{z}} = [\underline{z}_1, \dots, \underline{z}_k, \dots, \underline{z}_m]^{\mathrm{T}}$ be a multivariate normal vector with correlation matrix (equivalent):

$$\widetilde{\boldsymbol{C}}_{\underline{\boldsymbol{z}}} = \begin{pmatrix} 1 & \cdots & \tilde{\rho}_{1,m} \\ \vdots & \ddots & \vdots \\ \tilde{\rho}_{m,1} & \cdots & 1 \end{pmatrix}$$

In order to obtain $\boldsymbol{\underline{x}}$ through $\boldsymbol{\underline{z}}$ the following mapping equation is employed:

$$\underline{x}_k = F_{\underline{x}_k}^{-1}[\Phi(\underline{z}_k)] \tag{6}$$

where $F_{\underline{x}_k}^{-1}$ is the ICDF of variable $k$ and $\Phi(\cdot)$ is the standard normal CDF. A direct outcome of Eq. (6) is that for two variables $\underline{x}_k$ and $\underline{x}_l$ their correlation is given by:

$$\mathrm{Corr}[\underline{x}_k, \underline{x}_l] = \rho_{k,l} = \mathrm{Corr}\left[F_{\underline{x}_k}^{-1}[\Phi(\underline{z}_k)], F_{\underline{x}_l}^{-1}[\Phi(\underline{z}_l)]\right] \tag{7}$$

thus the target correlations $\rho_{k,l}$ are associated with the unknowns $\tilde{\rho}_{k,l}$.

An apparent approach could be setting $\widetilde{\boldsymbol{C}}_{\underline{\boldsymbol{z}}} \equiv \boldsymbol{C}_{\underline{\boldsymbol{x}}}$ However, both theoretical and empirical evidence have indicated that this assumption will result to systematically underestimated correlations within the synthetic data. The theoretical justification of this behavior stems from the Pearson correlation coefficient itself, since it is not invariant under non-linear monotonic transformations, such as those imposed by the ICDFs [*Embrechts et al.*, 1999 p. 8]. More specifically, the largest the departure of the actual distribution, $F_{\underline{x}_k}$, from the normal one, the largest will be the underestimation. Therefore, and except the trivial normal case, in order to eliminate biases, we should assign *a priori* larger values to $\tilde{\rho}_{k,l}$.

Hopefully, NDM and its theoretical background can provide a theoretical solution to the above problem by means of an appropriate correlation matrix $\widetilde{\boldsymbol{C}}_{\underline{\boldsymbol{z}}}$ that leads to the target correlation matrix $\boldsymbol{C}_{\underline{\boldsymbol{x}}}$. As highlighted by *Liu and Der Kiureghian* [1986], in order to employ NDM it is essential to ensure 1) one to one mapping of Eq. (6), and 2) positive definite correlation matrix $\widetilde{\boldsymbol{C}}_{\underline{\boldsymbol{z}}}$. The former requirement is by definition valid in typical case of continuous distributions used in hydrology, while the latter is also usually satisfied, since the distances $\varepsilon_{k,l} \coloneqq |\rho_{k,l} - \tilde{\rho}_{k,l}|$ are expected to be generally small (provided, of course, that the target matrix $\boldsymbol{C}_{\underline{\boldsymbol{x}}}$ is positive definite).

The following procedure is applied to each specific pair of variables $\underline{x}_k$ and $\underline{x}_l$ (i.e., $m(m-1)/2$ times). Given that

$$\mathrm{Corr}[\underline{x}_k, \underline{x}_l] = \rho_{k,l} = \frac{\mathrm{E}[\underline{x}_k, \underline{x}_l] - \mathrm{E}[\underline{x}_k]E[\underline{x}_l]}{\sqrt{\mathrm{Var}[\underline{x}_k]\mathrm{Var}[\underline{x}_l]}} \tag{8}$$

where $\mathrm{E}[\underline{x}_k], \mathrm{E}[\underline{x}_l]$ and $\mathrm{Var}[\underline{x}_k], \mathrm{Var}[\underline{x}_l]$ are the mean and variance of $\underline{x}_k$ and $\underline{x}_l$ respectively, which are obviously known since the associated marginal distributions are already specified (and have finite moments, otherwise the Pearson correlation coefficient cannot be defined) the computational procedure is limited to identifying $\mathrm{E}[\underline{x}_k, \underline{x}_l]$. Since the corresponding variables to be mapped, $\underline{z}_k$ and $\underline{z}_l$ , respectively, are by definition normally distributed, with correlation $\mathrm{Corr}[\underline{z}_k, \underline{z}_l] = \tilde{\rho}_{k,l}$, then, using (6) and the first cross-product moment of $\underline{x}_k$ and $\underline{x}_l$ we get:

$$\mathrm{E}[\underline{x}_k, \underline{x}_l] = \mathrm{E}\left[F_{\underline{x}_k}^{-1}[\Phi(\underline{z}_k)]F_{\underline{x}_l}^{-1}[\Phi(\underline{z}_l)]\right]$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{\underline{x}_k}^{-1}[\Phi(z_k)]F_{\underline{x}_l}^{-1}[\Phi(z_l)]\, \varphi_2(z_k, z_l, \tilde{\rho}_{k,l}) dz_k dz_l \tag{9}$$

where $\varphi_2(z_k, z_l, \tilde{\rho}_{k,l})$ is the bivariate standard normal probability density function. Therefore, each target $\rho_{k,l}$ is a function of $\tilde{\rho}_{k,l}$, which is embedded in $\varphi_2(z_k, z_l, \tilde{\rho}_{k,l})$, and the given marginal distributions $F_{\underline{x}_k}$ and $F_{\underline{x}_l}$, i.e.,

$$\rho_{k,l} = \mathcal{F}(\tilde{\rho}_{k,l}|F_{\underline{x}_k}, F_{\underline{x}_l}). \tag{10}$$

Unfortunately, Eq. (10) cannot be analytically derived from Eq. (9), with the exception of few special cases [*Li and Hammond*, 1975; *Cario and Nelson*, 1997]. Among them the Log-Normal case [*Mostafa and Mahmoud*, 1964] which is of particular interest in hydrology. The aforementioned researchers, as well as *Liu and Der Kiureghian* [1986], provided several Lemmas that can be useful in order to approximate Eq. (10). Among them,

**Lemma 1:** $\rho_{k,l}$ is a strictly increasing function of $\tilde{\rho}_{k,l}$.

**Lemma 2:** $\tilde{\rho}_{k,l} = 0$ for $\rho_{k,l} = 0$ as well as, $\tilde{\rho}_{k,l} \geq (\leq)\ 0$ if $\rho_{k,l} \geq (\leq)\ 0$.

**Lemma 3:** $|\rho_{k,l}| \leq |\tilde{\rho}_{k,l}|$.

Note that in Lemma 3, the equality sign is valid when $\rho_{k,l} = 0$ or when both marginal distributions are normal. Furthermore, the minimum and maximum attainable value of $\rho_{k,l}$ is given for $\tilde{\rho}_{k,l} = -1$ and $\tilde{\rho}_{k,l} = 1$, respectively. The literature offers a variety of approaches to establish $\mathcal{F}(\cdot)$, including crude search procedures [*Cario and Nelson*, 1996], methods based on the Gauss-Kronrod quadrature rule [*Cario*, 1996], root finding methods [*Li and Hammond*, 1975; *Chen*, 2001] as well as Gauss–Hermite quadrature and Monte-Carlo methods [*Xiao*, 2014]. Herein, we propose a simple and easy to implement method based on hybrid combination of Monte-Carlo simulation and polynomial interpolation.

### 4.3 Hybrid procedure for solving $\mathcal{F}(\cdot)$

As already mentioned, in order to preserve the target correlations $\rho_{k,l}$ in the actual domain, after mapping the generated Gaussian values with their prescribed distributions, using Eq. (6), it is essential to establish a suitable relationship between $\tilde{\rho}_{k,l}$ and $\rho_{k,l}$ . In this context, we have developed the following procedure (indices *k* and *l* are omitted for simplicity):

**Step 1:** Create a $q$-dimensional vector $\tilde{\boldsymbol{r}} = [\tilde{r}^1, \dots, \tilde{r}^q]$ of equally spaced values in the interval $[r_{\min}, r_{\max}]$. Here, lemma 2 can be accounted for in order to determine the boundaries $r_{\min}$ and $r_{\max}$, since it provides insights regarding the sign of $\tilde{\rho}$. For example, if the target correlation $\rho$ is positive, then we set $r_{\min} = 0$ and $r_{\max} = 1$.

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

**Step 3:** Map the synthetic data to the actual domain through Eq. (6), using the associated target marginal distribution,

**Step 4:** Calculate the empirical correlations $r^i$ and store them in the vector $\boldsymbol{r} = [r^1, \dots, r^q]$.

**Step 5:** Approximate the relationship between target ($\rho$) and equivalent ($\tilde{\rho}$) correlation by establishing a polynomial function of order *p*, among the values of $\tilde{\boldsymbol{r}}$ and $\boldsymbol{r}$ i.e.:

$$\rho = \mathcal{F}\left(\tilde{\rho} \middle| F_{\underline{x}_k}, F_{\underline{x}_l}\right) \cong r = a_p \tilde{r}^p + a_{p-1} \tilde{r}^{p-1} + \dots + a_1 \tilde{r}^1 + a_0 \tag{11}$$

**Step 6:** Evaluate the equivalent correlation $\tilde{\rho}_{k,l}$ by inverting the relationship between the fitted polynomial and the target correlation $\rho_{k,l}$.

We highlight that, according to Weierstrass approximation theorem, the formulation of the polynomial expression (11) is theoretically feasible, since $\mathcal{F}(\cdot)$ is continuous and $\tilde{r}$ is bounded on the interval [-1, 1]. Moreover, we remark that the constant term $a_0$ could be omitted, as indicated by Lemma 2.

The above procedure, which is hybrid combination of Monte Carlo simulation and numerical interpolation through polynomial regression, uses three input arguments, i.e., the vector dimension *q*, the sample size *N*, and the polynomial order *p*. The first two influence the accuracy and computational effort of the Monte Carlo procedure, while the third influences the accuracy of the interpolation approach. Preliminary analysis detected that a good balance between accuracy and computational efficiency is ensured for *q* around 10 — 20, and *N* around 50 000 — 100 000 trials. Regarding the polynomial order, *Xiao* [2014] conducted an extensive analysis, with distributions exhibiting a wide range of skewness and kurtosis coefficients, and concluded that $\mathcal{F}(\cdot)$ can be accurately approximated by a polynomial of less than ninth degree ($p \leq 9$). Apparently, for $p = q - 1$, the polynomial passes exactly through all simulated points, yet, in order to ensure parsimony, it may be preferable employing a less complicated expression. In this vein, in order to avoid over-fitting, we propose adjusting the order of the polynomial with the use of cross-validation techniques or the Akaike information criterion [*Akaike*, 1974]. We note that in the basis of a systematic study one could identify alternative functions instead of polynomials in order to describe the relationship $\mathcal{F}(\cdot)$.

The key advantage of the proposed methodology, which is applicable for continuous, discrete or mixed-type distributions, is its simplicity and the fact that it doesn't depend on specialized algorithms to solve the double integral of Eq. (9), in order to obtain a valid expression $\mathcal{F}(\cdot)$. It is noteworthy that despite the iterative nature of the algorithm, its implementation in high-level programming languages, such as R or MATLAB, requires less than 1 second (assuming $N$ = 150 000 and $m$ = 20) on a modest 3.0 GHz Intel Dual-Core i5 processor with 4 GB RAM.

### 4.4 Numerical example

Consider a bivariate example with variables $\underline{x}_1$ and $\underline{x}_2$, representing either the same process at two adjacent seasons or two simultaneous processes at the same season. We assume that the two variables follow the same target marginal distribution ($F_{\underline{x}_1} \equiv F_{\underline{x}_2}$), i.e., the Pearson type-III, with probability density function:

$$f(x|a,b,c) = \frac{1}{|b|\Gamma(a)}\left(\frac{x-c}{b}\right)^{a-1} \exp\left(-\frac{x-c}{b}\right) \tag{12}$$

where $\Gamma(\cdot)$ is the gamma function, $a$, $b$ and $c$ are the shape, scale and location parameters respectively. For both variables, we assume the common parameter values $a$ = 1, $b$ = 20 and $c$ = 0.60, as well as a target correlation $\rho_{1,2} = 0.70$. Employing the NDM approach and the numerical method of section 4.3, with $q$ = 20, $N$ = 50 000 and $p$ = 2, we approximate $\mathcal{F}(\cdot)$ through the following polynomial (quadratic) function (indices are omitted for simplicity):

$$\rho = \mathcal{F}\left(\tilde{\rho}|F_{\underline{x}_1}, F_{\underline{x}_2}\right) \cong 0.2049\tilde{\rho}^2 + 0.7963\tilde{\rho} - 0.0009 \tag{13}$$

Given the relationship (13), it is easy to solve for the equivalent correlation $\tilde{\rho}_{1,2}$ which can be used for the generation of standard normal variables, $\underline{z}_1$ and $\underline{z}_2$, that results to the target value $\rho_{1,2}$. In particular, for $\rho_{1,2} = 0.7$ and inverting (13), we get $\tilde{\rho}_{1,2} = 0.739$.

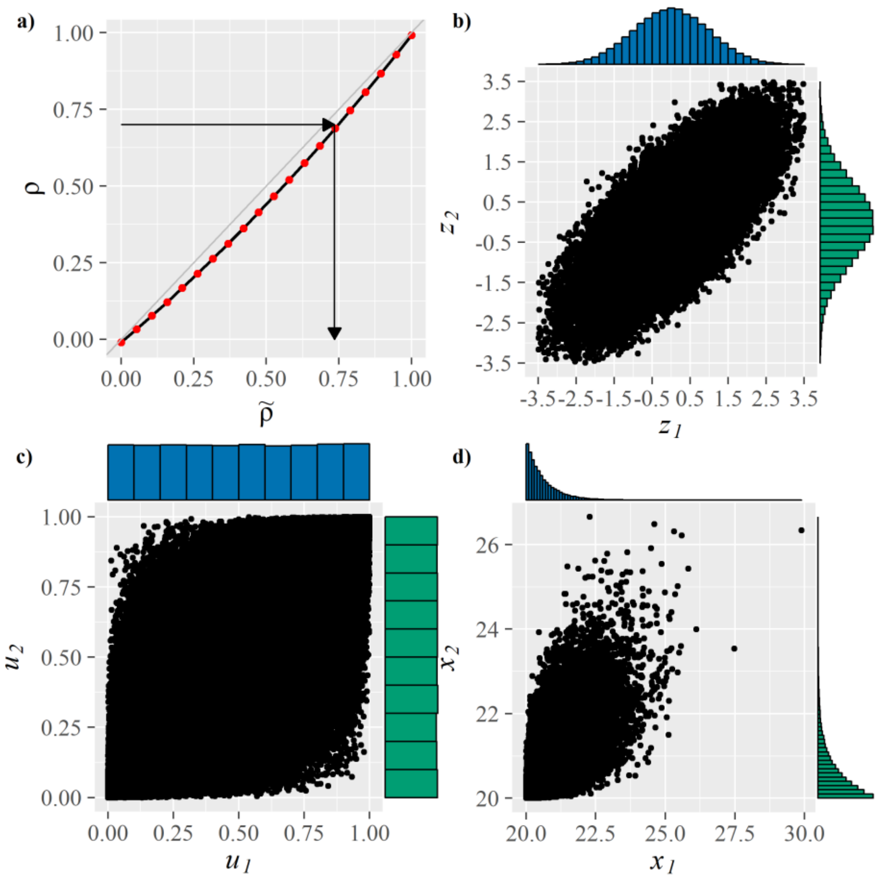


**Figure 1:** Hypothetical example of mapping two correlated variables, where the target and equivalent correlations represented through Eq. (13) are shown in panel (a). Panels (b), (c) and (d) illustrate the data in the normal, uniform and actual domain, respectively.

The mapping procedure of the numerical example, is shown in Figure 1 for the generation of 100 000 correlated values. In panel (a) we depict the relationship between target and equivalent correlations as established via Eq. (13). In panel (b) we illustrate the simulated auxiliary Gaussian variables, $\underline{z}_1$ and $\underline{z}_2$, which are generated by assigning the equivalent correlation $\tilde{\rho}_{1,2}$. Initially, these variables are mapped to the uniform domain through function $\Phi(\cdot)$ (panel c), and then they are mapped to the actual domain (panel d), via the corresponding inverse functions, $F_{\underline{x}_1}^{-1}$ and$F_{\underline{x}_2}^{-1}$. Within the two mapping procedures, the equivalent correlation $\tilde{\rho}_{1,2}$ is progressively decreased, down to the target value $\rho_{1,2}$.

We remark that due to the very large sample size, the empirical correlation between the auxiliary synthetic variables $\underline{z}_1$ and $\underline{z}_2$ coincides the theoretical one, i.e., $\tilde{\rho}_{1,2} = 0.739$, while the empirical correlation between the actual variables $\underline{x}_1$ and $\underline{x}_2$ is 0.707, thus practically identical to the target value $\rho_{1,2} = 0.70$. Moreover, the empirically estimated parameters of the derived distributions are *a* = 0.947, *b* = 20.001 and *c* = 0.622, for the synthetic variable $\underline{x}_1$ and *a* = 0.921, *b* = 20.000 and *c* = 0.671 for $\underline{x}_2$. The aforementioned values, which were computed through the maximum likelihood estimation method (MLE), are in agreement with the theoretical ones.

### 4.5 Coupling SPARTA and NDM

It is apparent that in order to align NDM with SPARTA, we just have to set $\underline{x}_k \coloneqq \underline{x}_s^i$ and $\underline{x}_l \coloneqq \underline{x}_{s-\tau}^j$ throughout equations (7) to (10), and approximate the required (by the auxiliary model) equivalent correlation coefficients $\tilde{\rho}_{s,s-\tau}^{i,j}$ of the target correlations $\rho_{s,s-\tau}^{i,j}$. For the estimation of the equivalent correlations across all processes and seasons, we also offer the aforementioned hybrid computational procedure to approximate the relationship of Eq. (10), i.e., $\mathcal{F}(\cdot)$.

### 4.6 Previous applications of NDM in hydrology

NDM-based approaches have been widely applied in industrial, financial and operations research applications, as indicated from the popularity of the original article by *Nataf* [1962] and the relevant publications [e.g., *Liu and Der Kiureghian*, 1986; *Cario and Nelson*, 1996, 1997; *Biller and Nelson*, 2003].

While hydrological community does not make direct reference to NDM and the associated models, such as NORTA, ARTA, VARTA, etc., it actually shares the same rationale, even from the geneses of hydrological stochastics. Loosely speaking, the core idea of NDM comprises the initiation from the Gaussian domain, with properly adjusted correlation coefficients, and then a mapping to the desirable domain.

In particular, *Matalas* [1967] has studied the effects of logarithmic transformations in the context of synthesizing log-normally distributed processes, concluding that the so far prevailing transformation approach failed to resemble the historical statistics. To reestablish consistency, he developed a framework based on the generation of normal processes, and provided a set of theoretical equations to estimate the statistical parameters (including adjusted correlation coefficients) in the Log-Normal domain. Later, *Klemeš and Borůvka* [1974] developed a generation scheme for gamma-distributed univariate first-order Markov chains, through a mapping procedure of Gaussian processes with the use of adjusted correlation coefficients. More recently, *Kelly and Krzysztofowicz* [1997] proposed and illustrated through several hydrology-related applications, a flexible bivariate distribution model, termed meta-Gaussian, which builds upon the bivariate standard normal distribution and the normal quantile transformation. Furthermore, *Wilks* [1998], in the context of his widely known weather generation model, has also employed a transformation procedure initiating from the standard Gaussian distribution, coupled with an empirical method to estimate the adjusted correlations for the simulation of multivariate daily precipitation with mixed exponential distributions. This seminal work has triggered the development of improved schemes, supporting more distributions and correlation structures. Detailed reviews are provided by *Wilks and Wilby* [1999] and *Ailliot et al.* [2015]. Additionally, running advances in stochastic hydrology are also in alignment with NDM. In particular, in a similar vein, *Serinaldi and Lombardo* [2017] proposed a fast procedure for autocorrelated univariate binary processes, while *Lee* [2017] introduced a simulation-based method for Gamma-distributed precipitation. Finally, *Papalexiou* [2017] proposes an elegant and unified overview for synthetic data generation using autoregressive models.

# 5 Case studies

## 5.1 Univariate simulation with common distribution models

The first case study involves the simulation of monthly flow of Nile River at Aswan dam, based on a historical dataset from March 1870 to December 1945 [*Hipel and McLeod*, 1994]. The flows are characterized by strong seasonality and high correlations across all subsequent months (Figure 2). In order to demonstrate the performance of SPARTA against PAR, we compare the outcomes of a stochastic simulation scenario of 2 000 years length, which has been used several times in the past for providing synthetic flows [e.g., *Koutsoyiannis et al.*, 2008]. Since PAR(1) is typically coupled with Pearson type-III distribution for white noise generation (referred to as PAR-PIII model), in order to conduct a fair and meaningful evaluation, within SPARTA we also set this distribution as target one for all months (referred to as SPARTA-PIII model). We remind that SPARTA explicitly accounts for the marginal distribution of each season, while PAR-PIII, similarly to most linear stochastic models, attempts to resemble the statistical characteristics via implicitly representing the marginal distributions into the innovation term. We note that the multivariate formulation of PAR-PIII of order 1 is given in Appendix A.

It is remarked that due to the use of Pearson type-III distribution, which allows for negative location parameters, the two models can produce negative values that would not be acceptable in a real-world hydrological study. A typical way to address this inconsistency within both models is the artificial truncation of all synthetic values to zero, which would yet introduce bias to the stochastic structure of the synthetic processes. However, among the two models, SPARTA also offers a much more rigorous alternative, since the data are generated via the corresponding ICDFs. The latter property enables fitting another positively bounded distribution model (e.g., Gamma, Log-Normal, etc.) to the observed data that explicitly prohibits the generation of negative values.

The two models are evaluated through visual inspection of simulated against observed values of their monthly statistical characteristics, in terms of calculated values of mean, $\mu$, standard deviation, $\sigma$, skewness coefficient, $C_s$, and lag-1 month-to-month correlation, $\rho_1$ (Figure 2), as well as in terms of their monthly marginal distributions (Figure 3). It is noted that the latter statistics were calculated after truncation of negative values. Except for the trivial case of means and standard deviations, which are perfectly reproduced by both models, for the skewness and month-to-month correlations, only SPARTA-PIII ensures full consistency with the target values across all seasons. In addition, SPARTA-PIII fits perfectly the target theoretical distribution models, which is a direct outcome of employing the inverse mapping, while PAR-PIII occasionally deviates from the target distributions, and particularly their tails (e.g., in February, March, April and May).

To further highlight the advantages of SPARTA over PAR-PIII, we also investigate the derived dependence forms, by focusing on the scatter plots of the 12 pairs of adjacent monthly data sets (Figure 4). Interestingly, PAR-PIII, although it preserves quite satisfactory the key statistical characteristics, including the observed coefficients of correlation, it fails to capture the full extent of the observed patterns, in contrast to SPARTA-PIII, which generates well-spread data pairs which are in compliance with the observations. In particular, in the scatter plots of pairs December – January, January – February, February – March and March – April, it is evident that PAR-PIII not only fails to capture the dependence patterns of the historical data, but also seems fails to produce synthetic pairs out of a lower boundary. Therefore, the synthetic dependencies are not in good agreement with the observed ones, although the correlation coefficients themselves are reproduced with high accuracy.

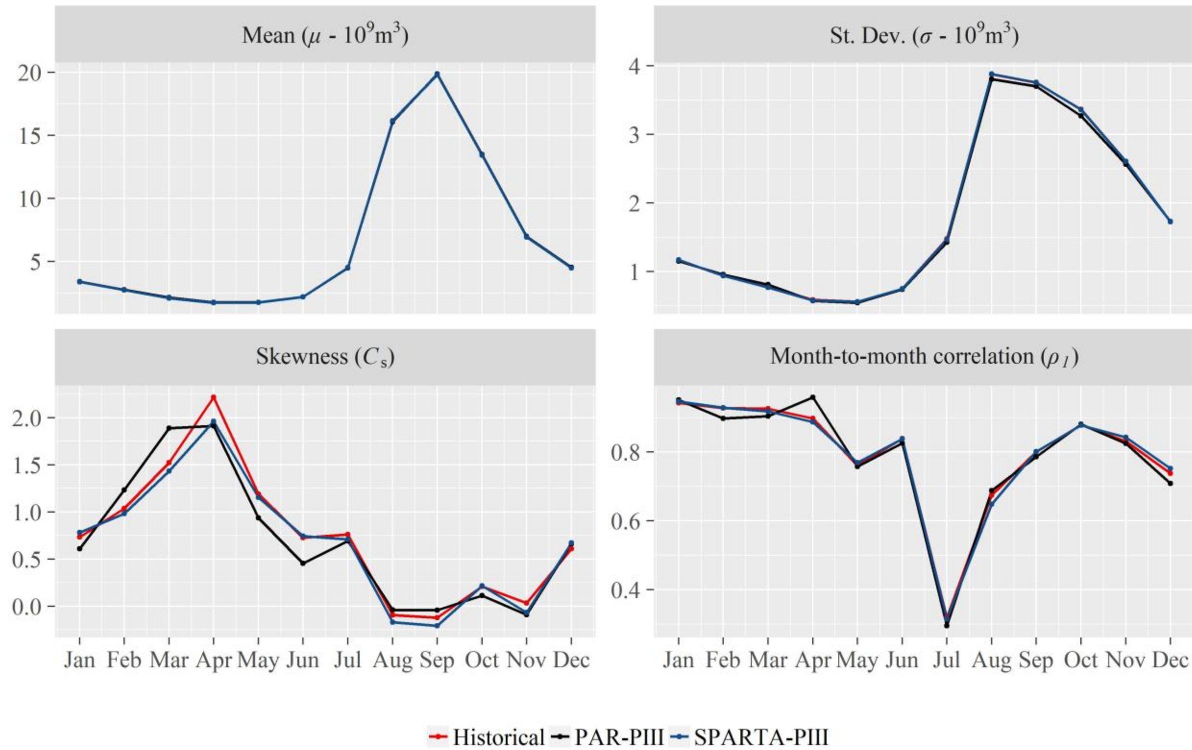


**Figure 2:** Comparison of key statistics ($\mu$, $\sigma$, $C_s$ and $\rho_1$) between historical and simulated flow data of Nile River (PAR and SPARTA).

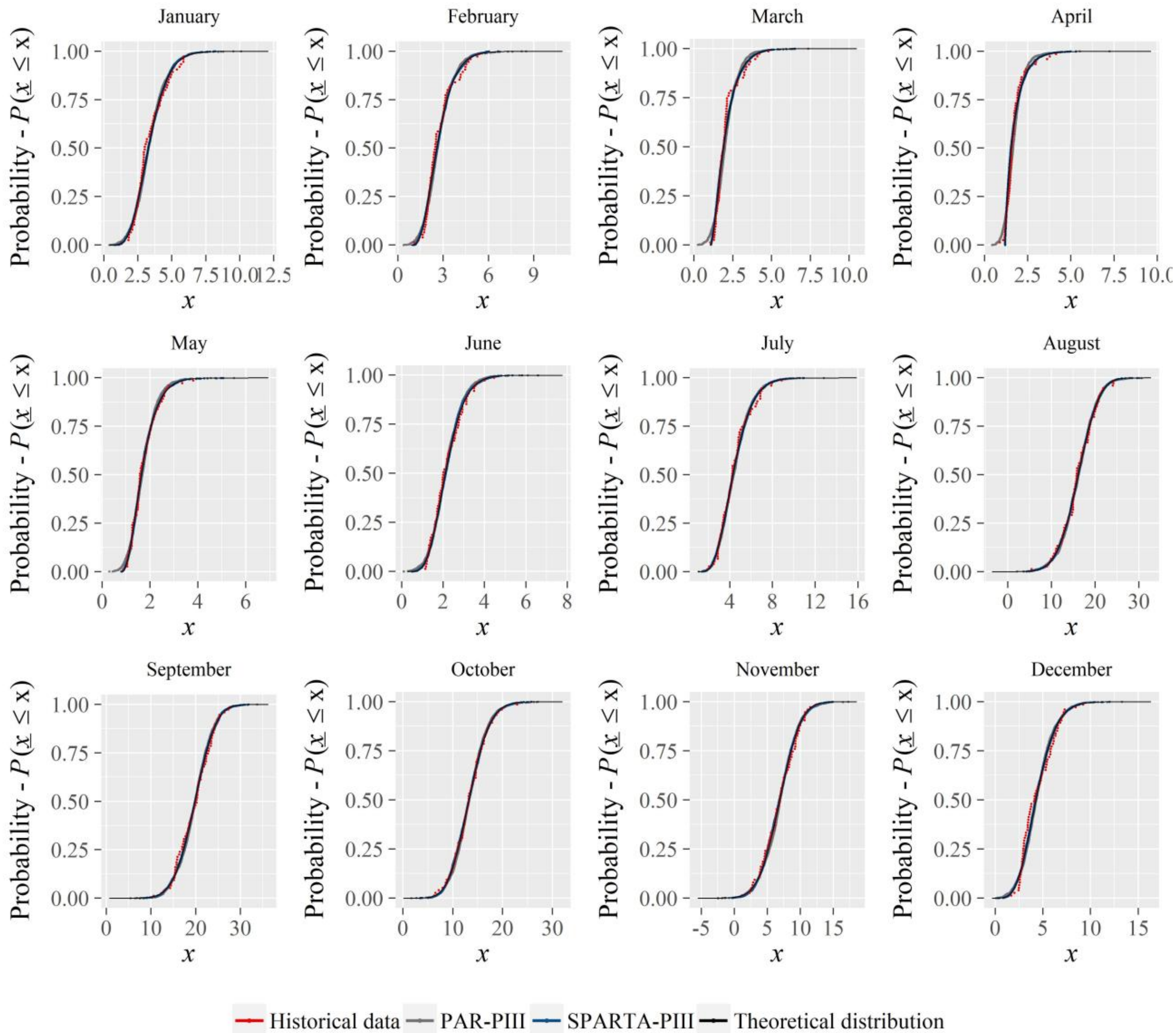


**Figure 3:** Comparison between simulated flow data ($10^9$ $m^3$), through PAR-PIII and SPARTA-PIII, empirical and theoretical cumulative distribution functions (Weibull plotting position). Simulated negative values are also included to avoid the distortion of the established CDFs.

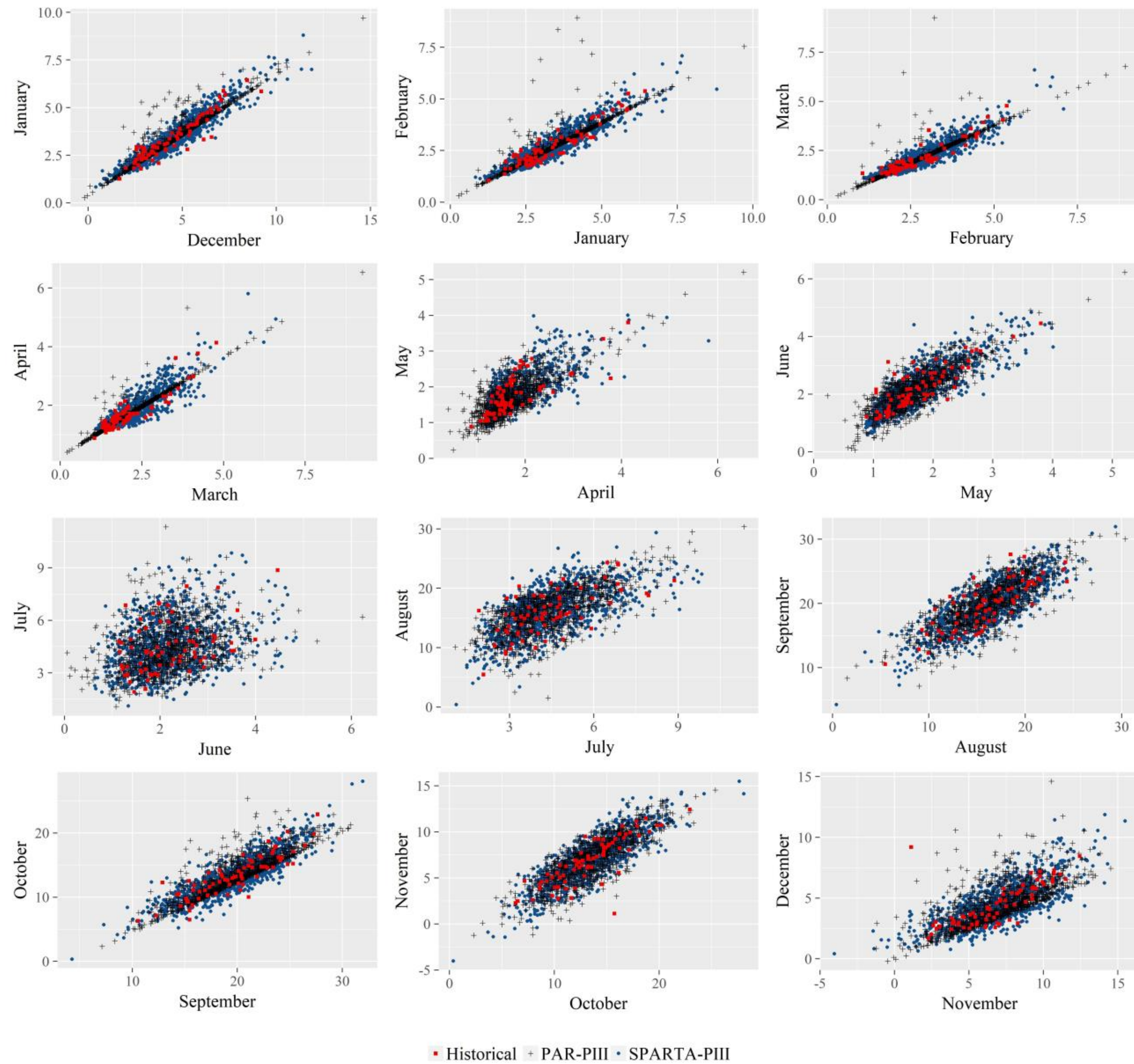


**Figure 4:** Month-to-month scatter plots of historical and simulated flow data ($10^9$ $m^3$), through PAR-PIII and SPARTA-PIII. Simulated negative values are also included to avoid the distortion of the established dependence patterns.

## 5.2 Toy simulation with seasonally-varying distribution models

The second case study involves the simulation of a hypothetical seasonal process, $\underline{x}_{s,t}$, with different marginal distribution per season (for convenience, 12 seasons are considered). The target distribution models and the associated parameters across seasons are given in Table 1. In addition, we assume the target lag-1 (i.e., season-to-season) correlation coefficients equal to $\boldsymbol{\rho} = [\rho_{12,1}, \rho_{1,2}, \dots, \rho_{s,s-1} \dots, \rho_{10,11}, \rho_{11,12}] = [0.93, 0.90, 0.76, 0.84, 0.32, 0.67, 0.80, 0.88, 0.83, 0.74, 0.94, 0.93]$. Using SPARTA we generated 1 000 × 12 = 12 000 synthetic values of $\underline{x}_{s,t}$ and compared their statistical characteristics against the target ones. We remark that in contrast to the previous case study, we do not compare against another linear stochastic model (e.g., PAR-PIII), given that we have specified different statistical distributions across seasons, which cannot be represented by such models.

The theoretical and simulated values of the key statistical characteristics of the modelled process are illustrated in Table 2. The former were calculated through the corresponding theoretical equations of each distribution. As shown, SPARTA is very efficient, since it reproduces all key statistics, including the kurtosis coefficient, $C_k$. Furthermore, SPARTA preserves the parameters of the target marginal distributions (Table 1, upper part), which are estimated through the MLE method. Actually, as shown in Table 1 (lower part), there is close agreement between the target and simulated parameter values for all seasons. This is also visually confirmed by plotting the associated CDFs (Figure 5), as the disparencies between the theoretical and empirical distributions are almost indistinguishable. It is noted that the distributions employed for season 5 and 10 allowed the generation of negative values since we assigned to the former a Gaussian one (which is unbounded) and in the latter a Pearson Type-III with location parameter $c$ = -50 which coincides with its theoretical lower bound (given that $b > 0$). All other distributions are defined in the positive real axis, hence they don't allow the generation of negative values.

Furthermore, the stochastic structure of the hypothetical process, by means of season-to-season correlations, $\rho_1$, is reproduced, despite the fact that it exhibits significant variability, also comprising some very high $\rho_1$ values. In order to shed further light on the seasonal dependence patterns, we provide scatter plots combined with histograms for four adjacent seasons, from which it becomes evident that SPARTA can reproduce a plethora of marginal distributions and simultaneously account for dependence patterns of different complexity (Figure 6).

**Table 1:** Theoretical distributions and associated parameters of hypothetical process across seasons, as well as MLE estimation of simulated data.

| Season | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Distribution/ Parameters | PIII | Exp | Gam | Norm | LoNo | Wei | Wei | LoNo | Exp | PIII | Wei | Gam |
| | | | | | | Theoretical Values | | | | | | |
| *a* | 1.7 | 0.015 | 10 | 85 | 5 | 4.5 | 6 | 6 | 0.003 | 11 | 3 | 9 |
| *b* | 10 | - | 0.15 | 30 | 0.3 | 680 | 820 | 0.25 | - | 19 | 155 | 0.2 |
| *c* | 40 | - | - | - | - | - | - | - | - | -50 | - | - |
| | | | | | | Simulated Values | | | | | | |
| *a* | 1.72 | 0.015 | 10.01 | 85 | 5 | 4.47 | 5.99 | 6 | 0.003 | 9.12 | 2.97 | 9.09 |
| *b* | 9.88 | - | 0.15 | 29.98 | 0.29 | 680.03 | 819.91 | 0.25 | - | 20.98 | 154.90 | 0.20 |
| *c* | 39.94 | - | - | - | - | - | - | - | - | -51.39 | - | - |

*Distribution abbreviations: PIII: Pearson type-III (*a* = shape, *b* = scale, *c* = location), Exp: Exponential (*a* = rate), Gam: Gamma (*a* = shape, *b* = rate), Norm: Normal (*a* = mean, *b* = st. dev.), LoNo: Log-Normal (*a* = log mean, *b* = log st. dev.), Wei: Weibull (*a* = shape, *b* = scale).

**Table 2:** Simulated and theoretical values of key statistical characteristics of hypothetical process.

| Season/ Statistic | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ (Theor.) | 57.00 | 66.67 | 66.67 | 85.00 | 155.24 | 620.55 | 760.72 | 416.23 | 333.33 | 159.00 | 138.41 | 45.00 |
| $\mu$ (Sim.) | 56.99 | 66.56 | 66.67 | 85.00 | 155.27 | 620.53 | 760.81 | 416.34 | 333.23 | 159.01 | 138.37 | 45.00 |
| $\sigma$ (Theor.) | 13.03 | 66.67 | 21.08 | 30.00 | 47.64 | 156.45 | 147.40 | 105.70 | 333.33 | 63.02 | 50.30 | 15.00 |
| $\sigma$ (Sim.) | 13.26 | 66.96 | 21.20 | 30.00 | 48.18 | 156.02 | 147.18 | 107.38 | 335.69 | 63.80 | 50.24 | 15.14 |
| $C_s$ (Theor.) | 1.53 | 2.00 | 0.63 | 0.00 | 0.97 | -0.17 | -0.37 | 0.88 | 2.00 | 0.60 | 0.16 | 0.66 |
| $C_s$ (Sim.) | 1.75 | 1.98 | 0.72 | -0.04 | 1.09 | -0.13 | -0.39 | 0.94 | 1.89 | 0.75 | 0.27 | 0.82 |
| $C_k$ (Theor.) | 6.53 | 9.00 | 3.60 | 3.00 | 4.99 | 2.80 | 3.03 | 4.06 | 9.00 | 3.54 | 2.72 | 3.66 |
| $C_k$ (Sim.) | 7.62 | 8.01 | 3.84 | 2.98 | 5.20 | 2.88 | 3.20 | 4.46 | 7.32 | 3.85 | 3.05 | 4.20 |
| $\rho_1$ (Theor.) | 0.93 | 0.90 | 0.76 | 0.84 | 0.32 | 0.67 | 0.80 | 0.88 | 0.83 | 0.74 | 0.94 | 0.93 |
| $\rho_1$ (Sim.) | 0.94 | 0.90 | 0.76 | 0.82 | 0.31 | 0.66 | 0.80 | 0.87 | 0.85 | 0.77 | 0.95 | 0.93 |
| $\tilde{\rho}_1$ (Equiv.) | 0.95 | 0.91 | 0.80 | 0.85 | 0.32 | 0.70 | 0.80 | 0.90 | 0.88 | 0.78 | 0.96 | 0.94 |

*Table abbreviations: Theor: Theoretical value, Sim: Simulated value, Equiv: Equivalent value.

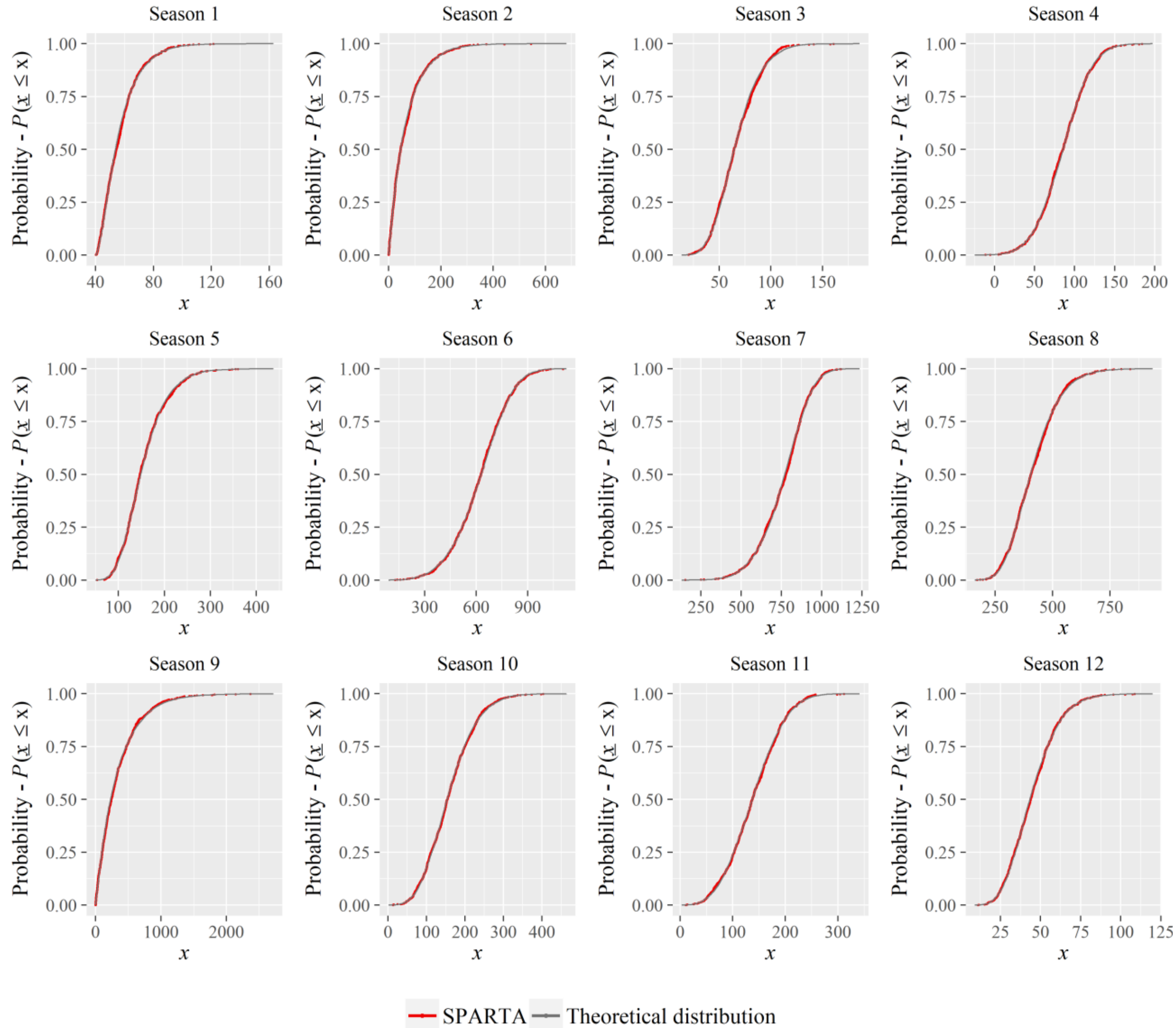


**Figure 5:** Comparison between simulated (SPARTA) and theoretical cumulative distribution functions (Weibull plotting position) of hypothetical process. Simulated negative values (season 5 and 10) are also included to avoid the distortion of the established CDFs.

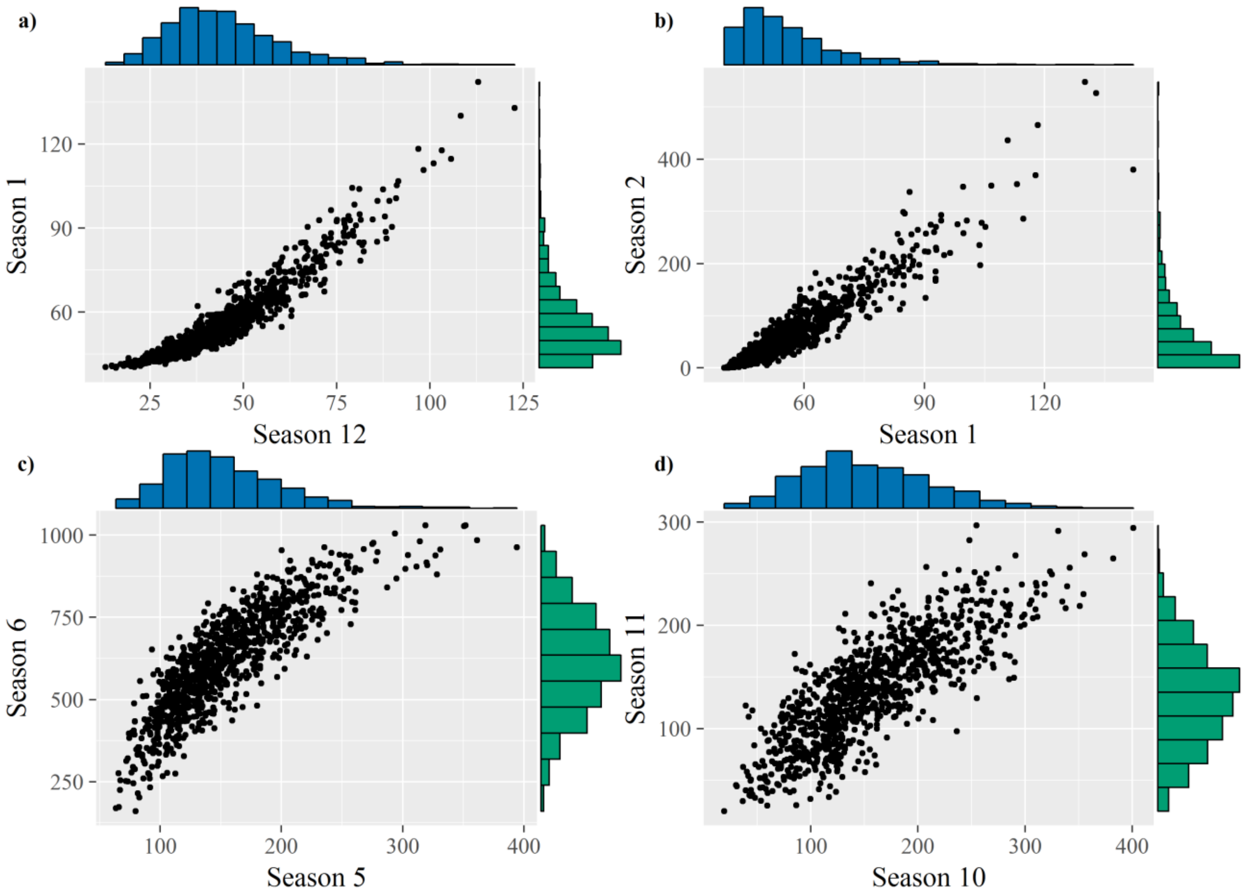


**Figure 6:** Scatter plots with histograms for a) season 12 vs. 1 b) season 1 vs. 2, c) season 5 vs. 6, and d) season 10 vs. 11.

### 5.3 Multivariate simulation

The third case study involves the simultaneous generation of monthly runoff and rainfall data at two major reservoirs of the water supply system of Athens, i.e., Evinos and Mornos (details about the system are provided by *Koutsoyiannis et al*, [2003]). The historical data cover a 29-year period (Oct/1979 – Sep/2008), which is marginally adequate for estimating up to third moment statistics with acceptable accuracy. For convenience, herein we will refer to Evinos runoff and rainfall as "sites" A and B, respectively, and to Mornos runoff and rainfall as "sites" C and D, respectively (here term "site" denotes a specific hydrological process at a specific location).

In this problem we employed the multivariate version of SPARTA and compared against the contemporaneous PAR(1) model with Pearson type-III white noise, again, referred as PAR-PIII model (Appendix A). Similarly to case study 1, in the context of specifying the underlying marginal distributions of SPARTA, and in order to ensure fair comparisons, we decided fitting the Pearson type-III model at all sites and for all months, and estimating its parameters via the method of moments. Under this premise, the generating scheme will be next referred to as SPARTA-PIII. Although we remark, that in an operational, "real-world study" one could take advantage of SPARTA model flexibility and select appropriate distributions models that are positively bounded, thus directly surpass the problem of negative values generation (see also the previous sections).

The performance of both models was assessed in a monthly basis, by contrasting the statistical characteristics of historical data that should be theoretically preserved by the corresponding generating schemes (i.e., monthly means, standard deviations, and skewness coefficients, lag-1 correlations across months, and zero-lag cross-correlations between all sites) against the simulated ones.

It is well-known that while the theoretical equations of any stochastic model are built in order to explicitly reproduce a specific set of statistical characteristics, this preservation is only ensured for very long (theoretically infinite) simulation horizons [*Efstratiadis et al.*, 2014]. If we consider relatively small horizons and repeat the simulation many times, the smaller the length of the synthetic sample, the larger is expected to be the variability of the simulated against the theoretical values of these characteristics. In this context, the stochastic model that ensures the minimum variability will be recognized as the most robust, since its performance will be the less sensitive against the simulation length. In this context, we employed two experiments, the first one by employing a single simulation of 500 000 years length, and the second one by running each model 500 times, to obtain independent synthetic samples of 1 000 years length. This Monte Carlo approach allowed for evaluating the uncertainty of the simulated statistical characteristics (after truncation of negative values to zero), which is depicted by means of box-plots (Figure 7 to Figure 11).

As shown in supplementary material (SM; Figure S1-S5), the estimated statistical characteristics from the large (i.e., 500 000 years) synthetic sample perfectly agree with the historical ones, thus confirming the solid theoretical background of SPARTA-PIII. As expected, PAR-PIII also ensures perfect fitting of the simulated to the observed statistics, expect for skewness, which are slightly underestimated. Probably, this systematic deviation is due to the simplified method employed for covariance matrix decompositions (namely, the Cholesky technique), as already mentioned in section 3.4.1.

The superiority of SPARTA-PIII against PAR-PIII is further revealed when evaluating the fitting of synthetic data to the theoretical distribution that has been adopted in this simulation experiment, i.e., Pearson type III. The latter is mathematically defined through Eq. (12) comprising three parameters, i.e., shape, $a$, scale, $b$, and location, $c$, which have been estimated for each site and each month with the method of moments (SM, Table S1). It is clearly shown that the estimated parameter values originated by SPARTA-PIII are very close to the theoretical ones, thus the desirable distributions are accurately reproduced. On the other hand, there are several cases where the PAR-derived parameters, and consequently the derived distributions, oscillate significantly form the theoretical model. This becomes even more evident when expressing these deviations in terms of root mean square error, per site and parameter. As shown in SM, Table S2, this error is up to three times larger than the error induced by SPARTA-PIII.

With respect to the second (i.e., Monte Carlo) experiment, from Figure 7 and Figure 8 it is shown that both SPARTA-PIII and PAR-PIII are able to reproduce the observed monthly means and standard deviations, respectively, since their variability is generally low across all sites and seasons. Regarding the reproduction of monthly coefficients of skewness (Figure 9), it seems that SPARTA-PIII slightly outperforms PAR-PIII in terms of statistical uncertainty, as indicated by the narrower box-plots that are provided is several cases (e.g., October, March, August and September for site A, October, November and March for site B, November, December and March for site C, and March, August and September for site D). Finally, in terms of lag-1 month-to-month and lag-0 cross-correlations, both schemes ensure robustness, as illustrated in Figure 10 and Figure 11, respectively.

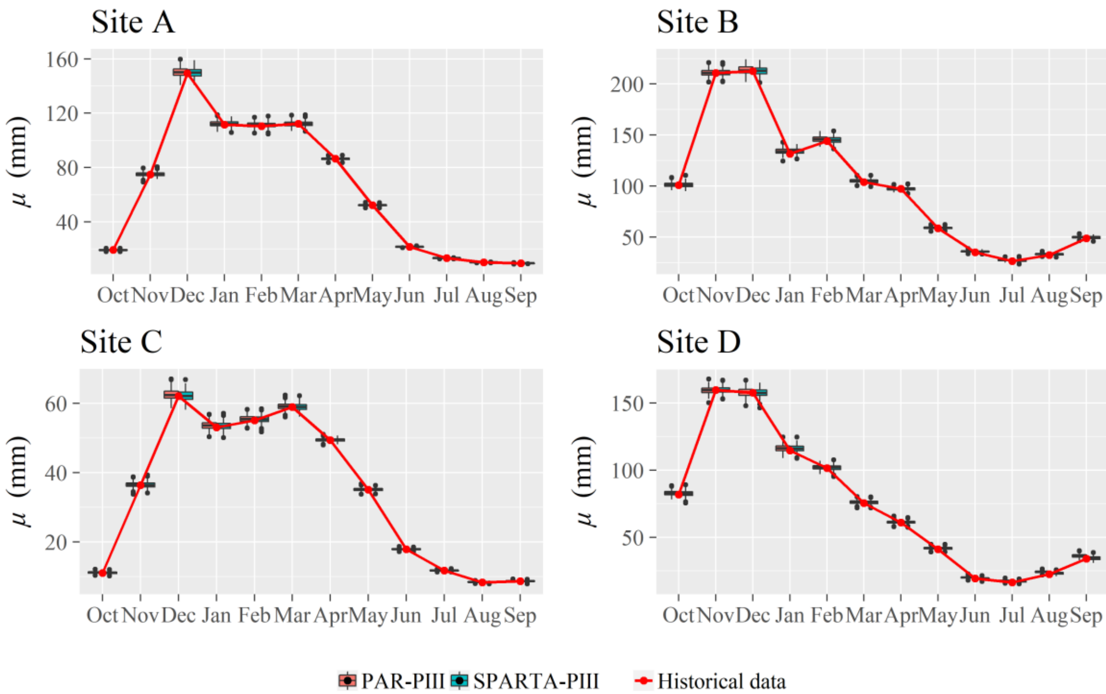


**Figure 7:** Comparison of monthly mean values, $\mu$, of historical and synthetic data.


**Figure 8:** Comparison of monthly standard deviation values, $\sigma$, of historical and synthetic data.

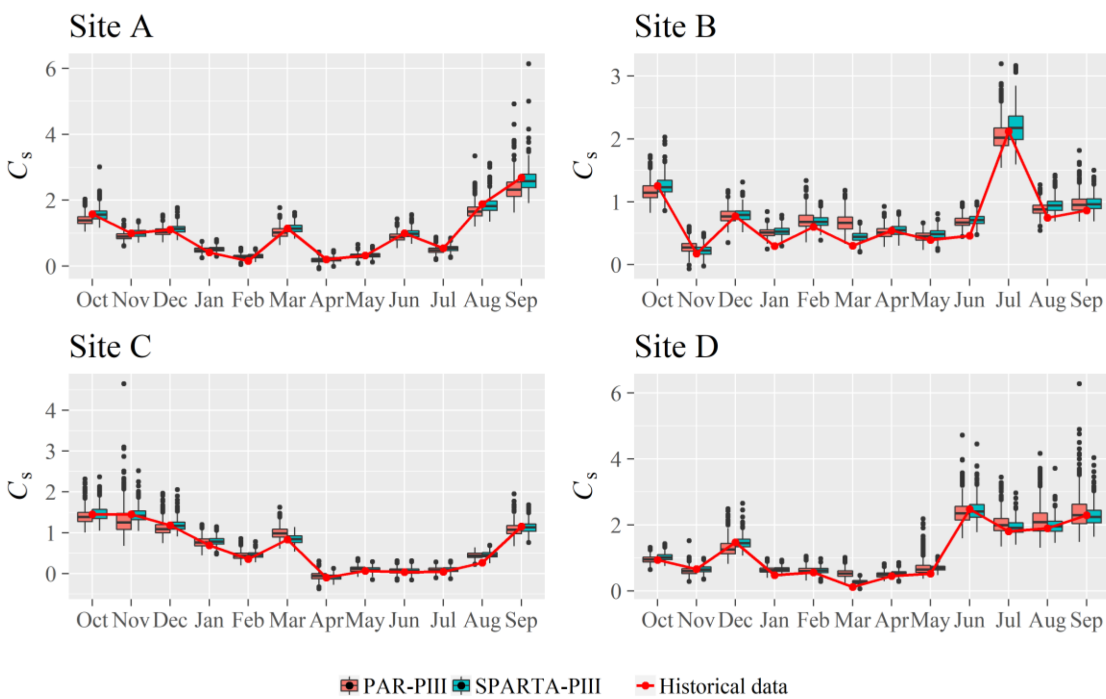


**Figure 9:** Comparison of monthly skewness coefficients, $C_s$, of historical and synthetic data.

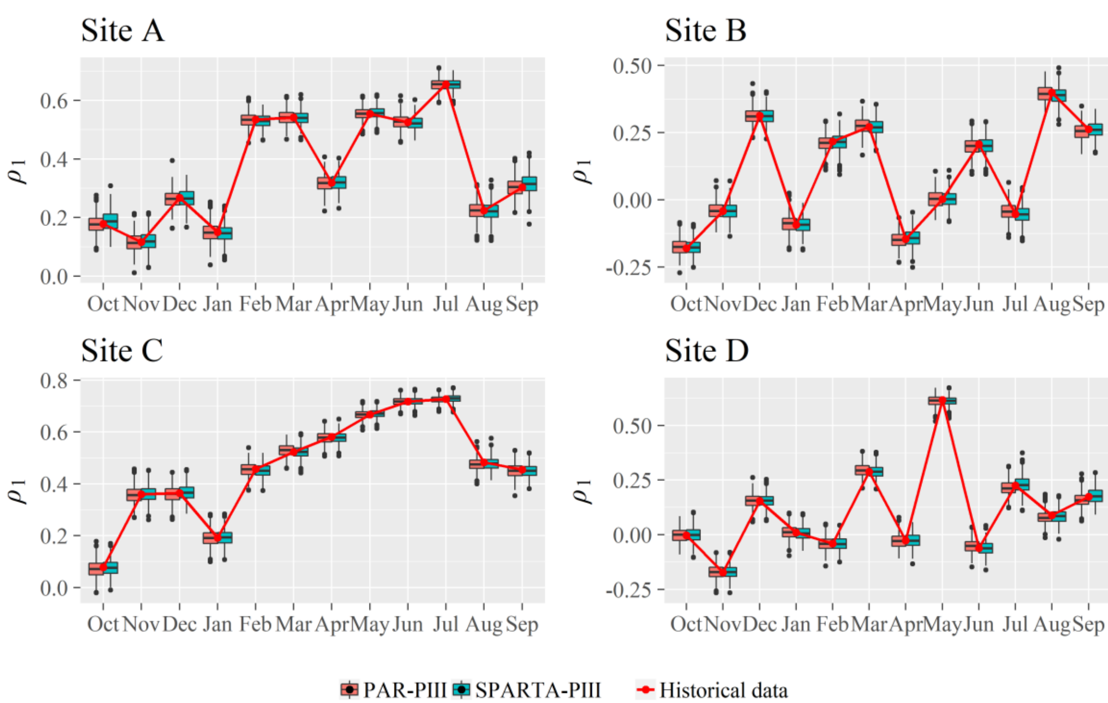


**Figure 10:** Comparison of month-to-month lag-1 correlations, $\rho_1$, of historical and synthetic data.

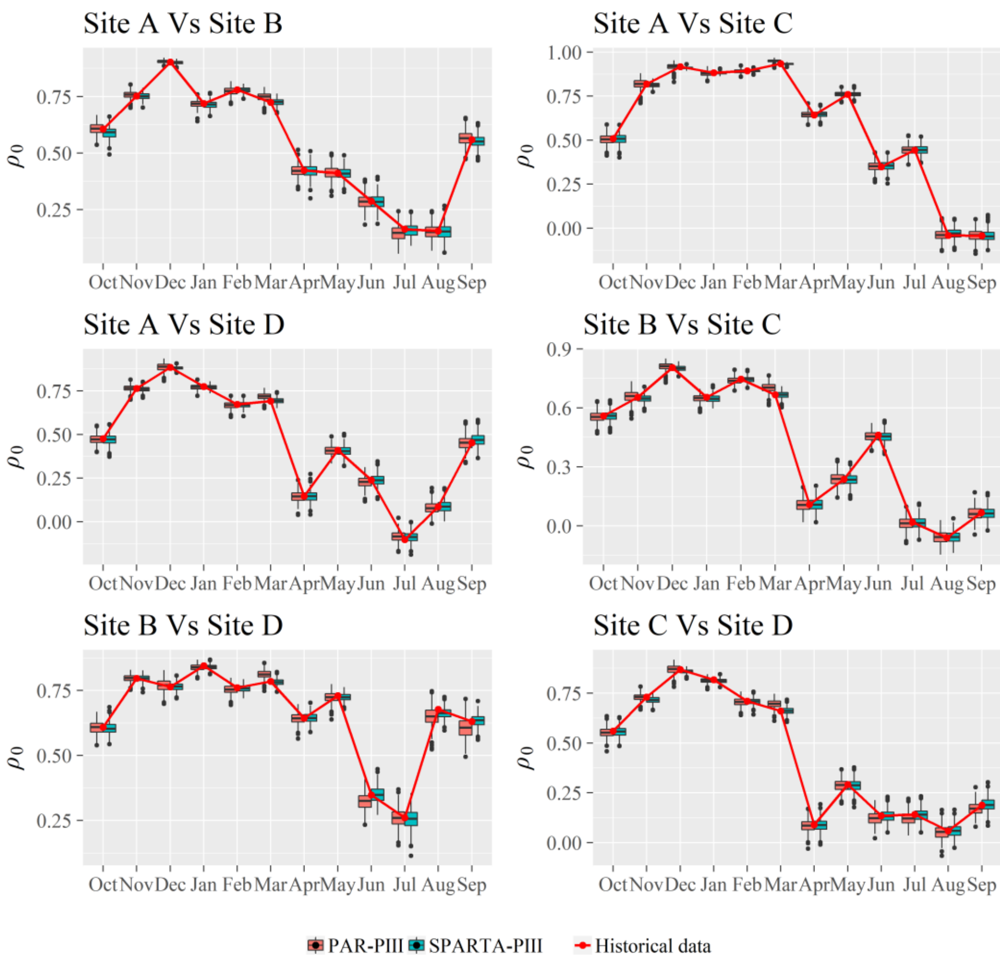


**Figure 11:** Comparison of monthly lag-0 cross-correlations, $\rho_0$, between sites of historical and synthetic data.

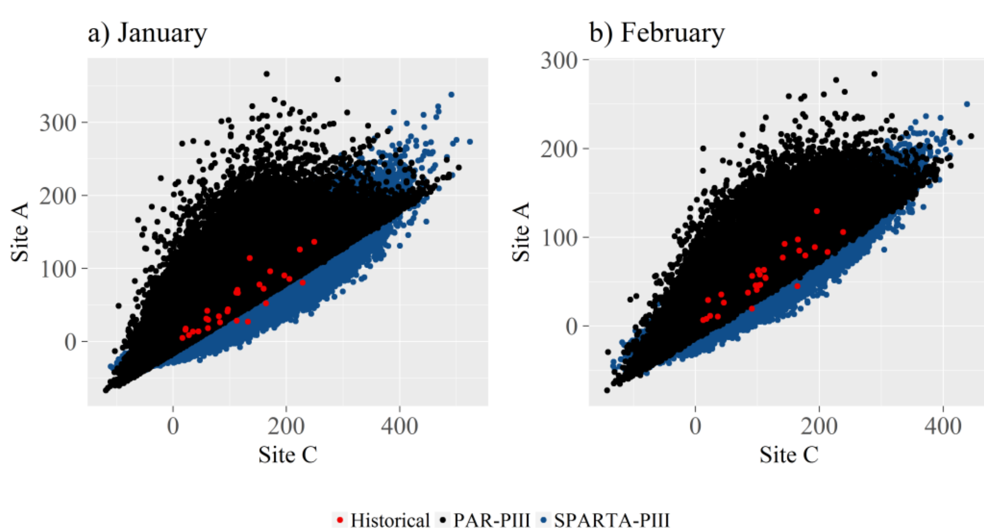


**Figure 12:** Scatter plots of 500 000 synthetic data for sites A and C, representing monthly runoff (mm) processes at Evinos and Mornos reservoirs, respectively, for (a) January and (b) February. Simulated negative values are also included to avoid the distortion of the established dependence patterns.

As already highlighted, a great advantage of SPARTA over linear stochastic schemes, such as PAR-PIII, is its ability to reproduce realistic dependence patterns, in compliance to the observed ones. This is also empirically confirmed in the current case study, which aims to reproduce both temporal and spatial dependencies (i.e., dependencies between different processes). A characteristic example is given in Figure 12, illustrating the scatter plots of historical and simulated runoff values of at Evinos (site A) and Mornos (site C), for months January and February, from the long-term experiment (i.e., 500 000 years). It becomes now even more evident that the SPARTA-PIII generation scheme provides reasonably-distributed data, while the synthetic data by PAR-PIII are again bounded within a specific range, which is far from truthful and does not capture the full extent of the observed scatter (notice the incompatibility between the synthetic series of PAR-PIII and the historical data in Figure 12).

# 6 Discussion

As briefly discussed in the introduction, and demonstrated through three case studies, the need for generic simulation schemes that allow producing synthetic data from multiple distributions primarily originates from the fact that the statistical behavior of many of hydroclimatic processes is not satisfactory captured by classical stochastic models. Such models cannot reproduce significant statistical aspects of the simulated processes (e.g., maxima and minima, associated with the tails of the distribution), although the "essential", low-order statistical characteristics of the parent data may be well-preserved.

However, to our opinion, the overall question is not just a technical issue, i.e., providing better stochastic models, but, in a more general context, revisiting the "essentials" of synthetic data. In particular, we suggest moving from the preservation of a specific set of statistical characteristics, which are exclusively inferred from the observed data, to the preservation of *a priori* specified theoretical distributions that are hypothesized to be consistent with the anticipated stochastic behavior of the underlying processes.

We recognize that the assignment of a specific distribution model for each modelled process is not a straightforward task, since the true distribution will always be unknown. Obviously, for a given data sample one can fit a plethora of distributions, combined with different parameter estimation procedures (e.g., classical moments, L-moments, maximum likelihood), and use typical statistical tests to assess the "optimal" scheme. Even for a given set of statistical characteristics, multiple distributions may be used. However, theoretical reasons, such as the central limit theorem and the principle of maximum entropy, may induce the selection of a different distribution, even when the latter is not so favored by the data [e.g., *Koutsoyiannis,* 2005; *Papalexiou and Koutsoyiannis,* 2012]. In any case, particularly when the historical samples are short or not so much reliable, the selection of the most suitable distribution may be supported by hydrological evidence. For instance, one may take advantage of the statistical behavior of the underlying processes in the broader area, as validated by large-scale regional studies [e.g., *Blum et al.*, 2017].

A final remark involves the treatment of historical data themselves. Actually, the observed statistics are subject to biases and uncertainties induced by their estimation from relatively short records (e.g., unreasonably high skewness values, due to outliers). Several times, the use of data as the sole means for extracting the statistical characteristics of the process of interest may also result to severe inconsistencies, such as negative autocorrelations that do not have physical meaning in hydrology [*Koutsoyiannis*, 2000]. Particularly, in the latter case it may be wise to follow the paradigm of the aforementioned author and fit a theoretical model on the empirically derived autocorrelation coefficients. Nevertheless, it may be preferable to assign, even manually, realistic values to the "suspicious" parameters rather than leave the model employing erroneous values. Moreover, due to changing environmental and hydroclimatic conditions, the statistical information contained in historical data may not be fully representative of the "projected" future conditions. In this context, aiming to explore the effects of change, several researchers suggest perturbing the values of the statistical characteristics to be reproduced within synthetic data [e.g., *Nazemi et al.*, 2013; *Borgomeo et al.*, 2015], which obviously imply employing parameters different than the data-driven ones. Nevertheless, wherever it is necessary to manually assign target input values, these have to be checked against both physical consistency and hydrological evidence. In this vein, we remark that NDM-based models (e.g., ARTA, VARTA and SPARTA) are able to synthesize data from any distribution hence allowing their straightforward use in such studies. This can be easily accomplished by changing the parameters of the distribution functions (even the distribution functions themselves) or the correlation structure of the process and subsequently investigate the effects of such changes to the system under study.

## 7 Conclusions

This work presents a novel approach, termed SPARTA, for the explicit stochastic simulation of univariate and multivariate cyclostationary (i.e., periodic) processes with arbitrary marginal distributions. SPARTA uses an auxiliary Gaussian PAR process with properly identified parameters, such as after its mapping to the actual domain through the ICDFs, it results to a process with the target correlation structure and *a priori* specified marginal distributions. Since the temporal and spatial dependencies are typically expressed by means of Pearson correlation coefficients, we focus on the identification of equivalent correlation coefficients of the auxiliary processes to be used in the Gaussian domain, in order to attain the target correlations in the actual domain. In this context, we use the Nataf joint distribution model, originated from statistical sciences for the generation of correlated random variables with prescribed distributions. Based on the theoretical background of NDM, we have developed a simple, yet efficient Monte-Carlo based approach that allows for identifying the equivalent correlation coefficients, $\tilde{\rho}$, with low computational effort.

Despite the obvious benefit of simulating processes with any marginal distributions, the proposed approach is also flexible in implementing any distribution fitting method, offered by recent advances in statistical sciences. This flexibility also offers the capability of explicitly ensuring the generation of non-negative values within simulations, through selecting appropriate distributions that are positively bounded. This very important potential, which is not offered by most of known stochastic schemes used in hydrology, is attributed to the use of the ICDF; if the latter is positively bounded, the generated values will be by definition non-negative.

The advantages of SPARTA in practice, i.e., in the context of generating monthly synthetic data, have been illustrated through three stochastic simulation studies, emphasizing different aspects of the proposed methodology. Furthermore, in two out of three studies, SPARTA has been contrasted to the well-established linear stochastic model PAR-PIII, i.e., PAR(1) with Pearson type-III white noise. The major outcomes of our analyses are:

- Both models reproduced almost perfectly the essential statistical characteristics of the simulated processes up to second order (means, standard deviations, lag-1 month-to-month correlations (i.e., autocorrelations), zero-lag cross-correlations);

- SPARTA was also able to preserve with high accuracy the third order statistics, expressed in terms of skewness coefficients, while in several cases PAR-PIII provided quite underestimated skewness, which varied significantly across independently generated synthetic samples;
- SPARTA was able not only to preserve the theoretical statistical characteristics of the observed data but also the parameters of the prescribed marginal distributions, which is in fact the primary goal of simulation (see discussion);
- SPARTA produced dependence structures in time and space that are in agreement with the observed patterns, while, in some cases, PAR-PIII provided rather irregular scatter patterns that were fragmented out of the observed ranges.

To this end, it is argued, that SPARTA is a convenient way to simulate cyclostationary processes, either univariate or multivariate, yet it should not be regarded as a *panacea* for all kind of simulation problems, since it inherits the characteristics of the auxiliary process from the periodic autoregressive family. In this context, it cannot preserve the statistical characteristics at aggregated time scales, e.g., annual, including long-range dependence (Hurst phenomenon). For this reason, future research involves the integration of SPARTA within a multi-scale stochastic framework, allowing us to reproduce the desirable distribution and desirable correlation structures at multiple time scales, and also reproduce the peculiarities of different scales. As shown in the literature, an effective and efficient way to address this is through disaggregation techniques. For instance, the hybrid Monte Carlo procedure by *Koutsoyiannis and Manetas* [1996], which has been successfully implemented within advanced simulation schemes [e.g., *Efstratiadis et al.*, 2014; *Kossieris et al.*, 2016], can be easily aligned with SPARTA to ensure statistical consistency across scales.

As a concluding remark, and following the discussion of section 6, the authors would like to highlight the fact that the blind use of stochastic models, with overconfidence on historical data, may create a distorted "reality", thus feeding operational hydrological and water management studies with inconsistent synthetic inputs. In this vein, we recommend to turn our efforts into the selection of the suitable distribution model, as well as the careful assessment of the sample statistics, with emphasis to high order moments and correlations that are prone to uncertainties. Therefore, the flexibility of the proposed approach contributes towards the establishment of a new paradigm in hydrological stochastics.

## Acknowledgments

The authors would like to thank the Associate Editor and the three anonymous reviewers, for their constructive comments, suggestions and critique, which helped providing a much improved manuscript. **Data availability**: Nile streamflow data at Aswan dam was retrieved from an external source (http://www.stats.uwo.ca/faculty/mcleod/epubs/mhsets/). The historical dataset (runoff and rainfall) of Evinos and Mornos reservoirs is available at: http://www.itia.ntua.gr/1746/. **Code availability**: The developed R scripts and functions that implement the SPARTA model are available upon request to the authors.

## References


Ailliot, P., D. Allard, V. Monbet, and P. Naveau (2015), Stochastic weather generators: an overview of weather type models, *J. la Société Française Stat.*, *156*(1), 101–113.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, *19*(6), 716–723.

Biller, B., and B. L. Nelson (2003), Modeling and generating multivariate time-series input processes using a vector autoregressive technique, *ACM Trans. Model. Comput. Simul.*, *13*(3), 211–237, doi:10.1145/937332.937333.

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

Borgomeo, E., C. L. Farmer, and J. W. Hall (2015), Numerical rivers: A synthetic streamflow generator for water resources vulnerability assessments, *Water Resour. Res.*, *51*(7), 5382–5405, doi:10.1002/2014WR016827.

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

Cario, M. C. (1996), Modeling and simulating time series input processes with ARTAFACTS and ARTAGEN, in *Proceedings of the 28th conference on Winter simulation - WSC '96*, edited by J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, pp. 207–213, ACM Press, New York, New York, USA.

Cario, M. C., and B. L. Nelson (1996), Autoregressive to anything: Time-series input processes for simulation, *Oper. Res. Lett.*, *19*(2), 51–58, doi:10.1016/0167-6377(96)00017-X.

Cario, M. C., and B. L. Nelson (1997), Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, *Ind. Eng.*, 1–19.

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

Chen, L., V. P. Singh, S. Guo, J. Zhou, and J. Zhang (2015), Copula-based method for multisite monthly and daily streamflow simulation, *J. Hydrol.*, *528*, 369–384, doi:10.1016/j.jhydrol.2015.05.018.

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

Embrechts, P., A. J. McNeil, and D. Straumann (1999), Correlation and Dependence in Risk Management: Properties and Pitfalls, in *Risk Management*, edited by M. A. H. Dempster, pp. 176–223, Cambridge University Press, Cambridge.

Fernandez, B., and J. D. Salas (1986), Periodic Gamma Autoregressive Processes for Operational Hydrology, *Water Resour. Res.*, *22*(10), 1385–1396.

Fiering, B., and B. Jackson (1971), *Synthetic Streamflows*, Water Resources Monograph, American Geophysical Union, Washington, D. C.

Fiering, M. B. (1964), Multivariate technique for synthetic hydrology, *J. Hydraul. Div.*, *90*(5), 43–60.

Hao, Z., and V. P. Singh (2013), Modeling multisite streamflow dependence with maximum entropy copula, *Water Resour. Res.*, *49*(10), 7139–7143, doi:10.1002/wrcr.20523.

Herman, J. D., H. B. Zeff, J. R. Lamontagne, P. M. Reed, and G. W. Characklis (2016), Synthetic Drought Scenario Generation to Support Bottom-Up Water Supply Vulnerability Assessments, *J. Water Resour. Plan. Manag.*, 4016050, doi:10.1061/(ASCE)WR.1943-5452.0000701.

Higham, N. J. (2002), Computing the nearest correlation matrix--a problem from finance, *IMA J. Numer. Anal.*, *22*(3), 329–343, doi:10.1093/imanum/22.3.329.

Hipel, K. W., and A. I. McLeod (1994), *Time series modelling of water resources and environmental systems*, Elsevier.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*, John Wiley, New York, NY, USA.

Kelly, K. S., and R. Krzysztofowicz (1997), A bivariate meta-Gaussian density for use in hydrology, *Stoch. Hydrol. Hydraul.*, *11*(1), 17–31, doi:10.1007/BF02428423.

Kirsch, B. R., G. W. Characklis, and H. B. Zeff (2013), Evaluating the Impact of Alternative Hydro-Climate Scenarios on Transfer Agreements: Practical Improvement for Generating Synthetic Streamflows, *J. Water Resour. Plan. Manag.*, *139*(4), 396–406, doi:10.1061/(ASCE)WR.1943-5452.0000287.

Klemeš, V., and L. Borůvka (1974), Simulation of Gamma-Distributed First-Order Markov Chain, *Water Resour. Res.*, *10*(1), 87–91, doi:10.1029/WR010i001p00087.

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

Koutsoyiannis, D. (1999), Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology, *Water Resour. Res.*, *35*(4), 1219–1229, doi:10.1029/1998WR900093.

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

Koutsoyiannis, D. (2005), Uncertainty, entropy, scaling and hydrological stochastics. 1. Marginal distributional properties of hydrological processes and state scaling / Incertitude, entropie, effet d'échelle et propriétés stochastiques hydrologiques. 1. Propriétés distributionnel, *Hydrol. Sci. J.*, *50*(3), 381–404, doi:10.1623/hysj.50.3.381.65031.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*(7), 2105–2117, doi:10.1029/96WR00488.

Koutsoyiannis, D., G. Karavokiros, A. Efstratiadis, N. Mamassis, A. Koukouvinos, and A. Christofides (2003), A decision support system for the management of the water resource system of Athens, *Phys. Chem. Earth, Parts A/B/C*, *28*(14–15), 599–609, doi:10.1016/S1474-7065(03)00106-2.

Koutsoyiannis, D., H. Yao, and A. Georgakakos (2008), Medium-range flow prediction for the

Nile: a comparison of stochastic and deterministic methods, *Hydrol. Sci. Journal-Journal Des Sci. Hydrol.*, *53*(1), 142–164, doi:10.1623/hysj.53.1.142.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*(3), 679–693, doi:10.1029/95WR02966.

Lawrance, A. J., and P. A. W. Lewis (1981), A new autoregressive time series model in exponential variables (NEAR (1)), *Adv. Appl. Probab.*, *13*(4), 826–845.

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

Lee, T. (2017), Multisite stochastic simulation of daily precipitation from copula modeling with a gamma marginal distribution, *Theor. Appl. Climatol.*, doi:10.1007/s00704-017-2147-0.

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

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

Maass, A., M. M. Hufschmidt, R. Dorfman, H. A. Thomas, S. A. Marglin, G. M. Fair, B. T. Bower, W. W. Reedy, D. F. Manzer, and M. P. Barnett (1962), *Design of water-resource systems*, Cambridge: Harvard University Press.

Marković, Đ., J. Plavšić, N. Ilich, and S. Ilić (2015), Non-parametric Stochastic Generation of Streamflow Series at Multiple Locations, *Water Resour. Manag.*, *29*(13), 4787–4801, doi:10.1007/s11269-015-1090-z.

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

Matalas, N. C. (1975), Developments in stochastic hydrology, *Rev. Geophys.*, *13*(3), 67, doi:10.1029/RG013i003p00067.

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

Mehrotra, R., R. Srikanthan, and A. Sharma (2006), A comparison of three stochastic multi-site precipitation occurrence generators, *J. Hydrol.*, *331*(1–2), 280–292, doi:10.1016/j.jhydrol.2006.05.016.

Mostafa, M. D., and M. W. Mahmoud (1964), On the Problem of Estimation for the Bivariate Lognormal Distribution, *Source Biometrika Biometrika Trust*, *51*(34), 522–527.

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

Papalexiou, S. M. (2017), A unified theory for exact stochastic modelling of univariate and multivariate processes with continuous, mixed type, or discrete marginal distributions and any correlation structure,

Papalexiou, S. M., and D. Koutsoyiannis (2012), Entropy based derivation of probability distributions: A case study to daily rainfall, *Adv. Water Resour.*, *45*, 51–57, doi:10.1016/j.advwatres.2011.11.007.

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

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

Salas, J. D. (1993), Analysis and modeling of hydrologic time series, in *Handbook of hydrology*, edited by D. R. Maidment, p. Ch. 19.1-19.72, Mc-Graw-Hill, Inc.

Salas, J. D., and G. G. S. Pegram (1977), A seasonal multivariate multilag autoregressive model in hydrology, in *Proc. Third Int. Symp. on Theoretical and Applied Hydrology, Colorado State Univ., Fort Collins, CO, USA.*

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

Salas, J. D., G. Q. Tabios, and P. Bartolini (1985), Approaches to multivariate modeling of water resources time series, *J. Am. Water Resour. Assoc.*, *21*(4), 683–708, doi:10.1111/j.1752-1688.1985.tb05383.x.

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

Sharma, A., D. G. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291–308, doi:10.1029/96WR02839.

Sklar, A. (1973), Random variables, joint distribution functions, and copulas, *Kybernetika*, *9*(6), 449–460.

Srinivas, V. V., and K. Srinivasan (2005), Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows, *J. Hydrol.*, *302*(1–4), 307–330, doi:10.1016/j.jhydrol.2004.07.011.

Thomas, H. A., and R. P. Burden (1963), *Operations research in water quality management*, HARVARD UNIV CAMBRIDGE MASS DIV OF ENGINEERING AND APPLIED PHYSICS.

Thomas, H. A., and M. B. Fiering (1962), Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, *Des. water Resour. Syst.*, 459–493.

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

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

Wilks, D. S., and R. L. Wilby (1999), The weather generation game: a review of stochastic weather models, *Prog. Phys. Geogr.*, *23*(3), 329–357, doi:10.1191/030913399666525256.

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

## Appendix A

We briefly present the contemporaneous PAR(1) model with Pearson type-III (3-parameter Gamma) white noise (referred as PAR-PIII), for multivariate simulation of monthly time series. The model is able to preserve the essential statistics (i.e., mean, variance and skewness coefficient) as well as the lag-1 month-to-month correlations (i.e., autocorrelations) and the lag-0 cross-correlations between locations. Following the notation of *Koutsoyiannis* [1999], let $\underline{\boldsymbol{x}}_s = \left[\underline{x}_s^1, \dots, \underline{x}_s^m\right]^{\mathrm{T}}$ be a vector which of $m$ stochastically dependent processes at season $s$. The generating scheme is:

$$\underline{\boldsymbol{x}}_s = \boldsymbol{A}_s \underline{\boldsymbol{x}}_{s-1} + \boldsymbol{B}_s \underline{\boldsymbol{w}}_s \tag{A.1}$$

where $\boldsymbol{A}_s, \boldsymbol{B}_s$ are seasonally-varying $m \times m$ parameter matrices and $\underline{\boldsymbol{w}}_s = \left[\underline{w}_s^1, \dots, \underline{w}_s^m\right]^{\mathrm{T}}$ is a vector of independent random variables generated from Pearson type-III distribution. The matrices $\boldsymbol{A}_s$ are calculated as follows:

$$\boldsymbol{A}_s = \operatorname{diag}\left(\frac{\operatorname{Cov}\left[\underline{\boldsymbol{x}}_s^1, \underline{\boldsymbol{x}}_{s-1}^1\right]}{\operatorname{Var}\left[\underline{\boldsymbol{x}}_{s-1}^1\right]}, \dots, \frac{\operatorname{Cov}\left[\underline{\boldsymbol{x}}_s^m, \underline{\boldsymbol{x}}_{s-1}^m\right]}{\operatorname{Var}\left[\underline{\boldsymbol{x}}_{s-1}^m\right]}\right) \tag{A.2}$$

while matrices $\boldsymbol{B}_s$ are given by:

$$\boldsymbol{B}_s \boldsymbol{B}_s^{\mathrm{T}} = \boldsymbol{G}_s \tag{A.3}$$

where

$$\boldsymbol{G}_s = \operatorname{Cov}\left[\underline{\boldsymbol{x}}_s, \underline{\boldsymbol{x}}_s\right] - \boldsymbol{A}_s \operatorname{Cov}[\underline{\boldsymbol{x}}_{s-1}, \underline{\boldsymbol{x}}_{s-1}]\, \boldsymbol{A}_s^{\mathrm{T}} \tag{A.4}$$

where $\operatorname{Cov}[\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\psi}}]$ denotes the covariance of vectors $\underline{\boldsymbol{\xi}}$ and $\underline{\boldsymbol{\psi}}$, i.e., $\operatorname{Cov}\left[\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\psi}}\right] = \mathrm{E}\left\{\left(\underline{\boldsymbol{\xi}} - \mathrm{E}[\underline{\boldsymbol{\xi}}]\right)\left(\underline{\boldsymbol{\psi}}^{\mathrm{T}} - \mathrm{E}[\underline{\boldsymbol{\psi}}]^{\mathrm{T}}\right)\right\}$. At each season $s$, the parameter matrix $\boldsymbol{B}_s$ can be estimated either through typical decomposition techniques (e.g., Cholesky or singular value decomposition) or numerically approximated, e.g., through optimization approaches [*Koutsoyiannis*, 1999; *Higham*, 2002].

Regarding the white noise vector $\underline{\boldsymbol{w}}_s$, its statistical structure is associated with the seasonal statistical characteristics of the parent process, through the following equations:

$$\mathrm{E}\left[\underline{\boldsymbol{w}}_s\right] = \boldsymbol{B}_s^{-1}\left\{\mathrm{E}\left[\underline{\boldsymbol{x}}_s\right] - \boldsymbol{A}_s \mathrm{E}\left[\underline{\boldsymbol{x}}_{s-1}\right]\right\} \tag{A.5}$$

$$\operatorname{Var}\left[\underline{\boldsymbol{w}}_s\right] = [1, \dots, 1]^{\mathrm{T}} \tag{A.6}$$

$$\mu_3\left[\underline{\boldsymbol{w}}_s\right] = (\boldsymbol{B}_s^{(3)})^{-1}\left\{\mu_3\left[\underline{\boldsymbol{x}}_s\right] - \boldsymbol{A}_s^{(3)}\, \mu_3\left[\underline{\boldsymbol{x}}_{s-1}\right]\right\} \tag{A.7}$$

where $\boldsymbol{B}_s^{(k)}$ is a matrix whose elements are raised to power $k$ while $\mu_3\left[\underline{\boldsymbol{w}}_s\right]$ and $\mu_3\left[\underline{\boldsymbol{x}}_s\right]$ are vectors that denote the third central moments of $\underline{\boldsymbol{w}}_s$ and $\underline{\boldsymbol{x}}_s$ respectively. The white noise is produced by a suitable random number generator, in particular the Pearson type-III distribution, which can explicitly preserve $\mathrm{E}\left[\underline{\boldsymbol{w}}_s\right]$, $\operatorname{Var}\left[\underline{\boldsymbol{w}}_s\right]$ and $\mu_3\left[\underline{\boldsymbol{w}}_s\right]$.