15th International Conference on Environmental Science and Technology
Rhodes, Greece, 31 August to 2 September 2017

Stochastic simulation of periodic processes with arbitrary marginal distributions

Tsoukalas I.*, Efstratiadis A. and Makropoulos C.
Department of Water Resources and Environmental Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80, Zographou, Greece

*corresponding author

e-mail: itsoukal@mail.ntua.gr

Abstract Stochastic simulation of hydrological processes has a key role in water resources planning and management due to its ability to incorporate hydrological uncertainty within decision-making. Due to seasonality, the statistical characteristics of such processes are considered periodic functions, thus implying the use of cyclo-stationary stochastic models, typically using a common statistical distribution. Yet, this may not be representative of the statistical structure of such processes across all seasons. In this context, we introduce a novel model suitable for the simulation of periodic processes with arbitrary marginal distributions, called Stochastic Periodic AutoRegressive To Anything (SPARTA). Apart from capturing the periodic correlation structure of the underlying processes, its major advantages are a) the accurate preservation of seasonally-varying marginal distributions; b) the explicit generation of non-negative values; and c) the parsimonious model structure. Finally, the performance of the model is demonstrated through a theoretical (artificial) case study.

Keywords: Stochastic simulation, periodic processes, hydrological processes, arbitrary marginal distributions

1. Introduction

Two common peculiarities of time series (especially in hydrological domain) are non-Gaussianity and periodicity, with the latter implying a periodic fluctuation of the marginal statistics of the underlying process as well as a periodic correlation structure. Characteristic examples of such processes are the monthly time series of precipitation and river flow discharge. Concerning the modelling of such time series, it is known that the classic cyclic standardization approach (Kottegoda, 1980; Salas, 1993) is not able to capture the seasonally varying autocorrelation coefficients due to the underlying assumption of stationarity. On the contrary, cyclostationarity (i.e., seasonally varying parameters), allows the variation of such properties and hence it consists a more appropriate modelling scheme. The first cyclostationary model is attributed to Thomas-Fiering (1962) who developed a Gaussian univariate periodic simulation model able to preserve the lag-1 correlation between successive seasons. The seminal work of Thomas-Fiering have led to a broader family of models, termed periodic autoregressive (PAR). The latter family of models have been extensively studied by many researchers including higher order and multivariate implementations (Bras and Rodríguez-Iturbe, 1985; Kottegoda, 1980; Salas, 1993).

Further to periodicity, non-Gaussianity is another typical characteristic of hydrological variables, commonly observed across (almost) all time-scales. This highlights the necessity to account for skewed, non-Gaussian distributions. Early attempts to simulate non-normal time series involved their transformation to Gaussian via a normalization function; such as Box-Cox and logarithmic transformation. Next, parameter estimation and simulation is performed on the normalized data and the final product is obtained via the inverse transformation (Salas et al., 1985). However, in most cases, such simple transformations are not adequate and many attempts have been made using *ad-hoc* functions involving typically 4-5 parameters (e.g., Koutsoyiannis et al., 2008). Hence, this procedure can be characterized as non-trivial and prone to subjectivity. Note, that even if a proper normalization function is identified, it is not ensured that the normalization – simulation – de-normalization procedure will preserve the desired statistics or the stochastic structure of the original variables (Bras and Rodríguez-Iturbe, 1985; Salas et al., 1985). The latter highlight that failure or ill-transformation of the data to Gaussian may lead to miss-specification of the marginal statistics and inevitably lead to miss-specified models.

Probably due to the aforementioned shortcomings, the literature has lean towards approaches that incorporate skewness within the model structure; i.e., via generating white noise from a specific, skewed, distribution (Fiering and Jackson, 1971). Extended reviews regarding such methods can be found in literature (Matalas and Wallis, 1976; Salas et al., 1985) which also includes approaches with white noise generated from the Pearson type-III distribution (e.g., Efstratiadis et al., 2014; Koutsoyiannis and Manetas, 1996). The two notable shortcomings of such approaches are a) the generation of negative values and b) that they provide just an approximation of the variable's marginal distribution since the "strict exactness" is lost due to the underlying generation mechanism (Koutsoyiannis and Manetas, 1996).

In order to address the aforementioned issues, we propose a method for generating periodic processes with

arbitrary marginal distributions while preserving simultaneously the stochastic structure of the processes. Our method, called Stochastic Periodic AutoRegressive To Anything (SPARTA, Tsoukalas et al., 2017) constitutes a generalization of the univariate AutoRegressive To Anything (ARTA) model of Cario and Nelson (1996) for periodic processes. The central idea involves a) the simulation of an auxiliary periodic PAR process upon the "Gaussian" domain with such parameters that capture the stochastic structure (season-to-season autocorrelation) of the process, and b) the mapping of the generated series to the "real" domain, via the inverse cumulative distribution function (ICDF). The main challenge encountered in the aforementioned methods is the identification of the parameters of the auxiliary process that result in the desired stochastic structure after the application of the inverse cumulative distribution function. This arises from the fact that Pearson correlation coefficient, which is used within the parameter identification procedure of both AR and PAR models, is not invariant under monotonic transformations; such as those imposed by the inverse of the desired distribution. Therefore, we have to identify the "equivalent" correlation coefficient that should be used within the parameter identification procedure of the auxiliary PAR model in order to attain the desired correlation after the mapping to the "real" domain. The estimation of "equivalent" correlation coefficient requires the integration of a double infinite integral which can be easily accomplished with the use of numerical methods. The latter joint relationship is known as Nataf distribution model (Nataf, 1962).

The main advantages of the proposed methodology are a) its ability to account for the cyclostationarity and simultaneously simulate time series with arbitrary marginal distributions b) the flexibility provided in the selection of distribution fitting method and c) the parsimonious model structure, since SPARTA uses exactly the same number of parameters as PAR model.

2. Methodology

The key idea behind SPARTA model lies in the simulation of an auxiliary univariate periodic Gaussian process $\{Z_s\}$; where s refers to season; with such parameters (which define the stochastic structure) that after the mapping with the corresponding inverse distribution function results into a process $\{X_s\}$ with the desired correlation structure and marginal distributions. The mapping operations is of the following form:

$$X_s = F_{X_s}^{-1}[\Phi(Z_s)] \qquad (1)$$

Where $\Phi(\cdot)$ refers to the standard normal cumulative distribution function (CDF) and $F_{X_s}^{-1}(\cdot)$ denotes the ICDF of the desired distribution. Briefly, the methodology can be summarized in five steps:

a) Define (i.e., fit) a suitable marginal distribution function F_{X_s}, to each season.

b) Select an appropriate auxiliary periodic Gaussian model (e.g., PAR(1)).

c) Approximate the equivalent correlation of pairs of interest (e.g., those related with the model parameters).

d) Estimate the parameters of the auxiliary process $\{Z_s\}$ using the equivalent correlations identified in step c.

e) Simulate a realization of the auxiliary process $\{Z_s\}$ and map the generated data to the real domain (using eq. (1)), in order to attain the process $\{X_s\}$, using the ICDFs identified in step a.

Although the proposed methodology is generic and higher order models can be employed, here we prefer to use the PAR(1) model in order to keep things simple and provide an easy to follow narrative. Furthermore, our choice regarding the PAR(1) model is further supported by the findings of other researchers that highlight that the parsimonious structure of PAR(1) model is adequate for the simulation of hydrological time series (e.g., Efstratiadis et al., 2014; Koutsoyiannis and Manetas, 1996). Therefore, prior to describing the methodology for the identification of the equivalent correlation allow us first to describe the auxiliary univariate PAR(1) model. Hereafter we will symbolize the equivalent correlation in Gaussian space as $\hat{\rho}_{(\cdot)}$ and the desired correlation in the real domain as $\rho_{(\cdot)}$. The key equation of the univariate PAR(1) model is of the form:

$$Z_s = \hat{\rho}_{s,s-1} Z_{s-1} + \sqrt{1 - \hat{\rho}_{s,s-1}{}^2}\, W_s \qquad (2)$$

Where W_s is an independent identically distributed variable from N ~(0, 1). It can be shown that the resulting process $\{Z_s\}$ will have marginal distributions N ~(0, 1), which in combination with eq. (1) ensures that the process $\{X_s\}$ will have the desired distribution.

Therefore, the main challenge of the aforementioned procedure is to identify the equivalent correlation coefficient $\hat{\rho}_{s,s-1}$ which should be used in the auxiliary process $\{Z_s\}$. For notational purposes allow us to define the following indices, $X_i \coloneqq X_s$ and $X_j \coloneqq X_{s-1}$. The season-to-season correlation structure of the $\{Z_s\}$ process is associated with that of $\{X_s\}$ since,

$$\rho_{i,j} = \mathrm{Corr}[X_i, X_j] = \mathrm{Corr}\left\{F_{X_i}^{-1}[\Phi(Z_i)], F_{X_j}^{-1}[\Phi(Z_j)]\right\}$$

for all $i \neq j$. As shown in Nataf (1962), as well as, in Cario and Nelson (1997) the latter relationship is limited to adjusting $E[X_i, X_j]$, since,

$$\mathrm{Corr}[X_i, X_j] = \rho_{i,j} = \frac{E[X_i, X_j] - \mu_i \mu_j}{\sigma_i \sigma_j} \qquad (3)$$

Where μ_i, μ_j and σ_i, σ_j denote the mean and the standard deviation of X_i and X_j respectively, which can be derived from corresponding marginal distributions. Then since the relationship between Z_i and Z_j is expressed via the bivariate standard normal distribution with correlation $\mathrm{Corr}[Z_i, Z_j] = \hat{\rho}_{i,j}$ and with the use of the first cross product moment of X_i and X_j we obtain the following equation,

$$E[X_i, X_j] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X_i}^{-1}[\Phi(z_i)]F_{X_j}^{-1}[\Phi(z_j)]\ \varphi(z_i, z_j, \hat{\rho}_{i,j})\mathrm{d}z_i\mathrm{d}z_j \quad (4)$$

Where $\varphi(z_i, z_j, \hat{\rho}_{i,j})$ is the bivariate normal probability density function (PDF) with correlation $\hat{\rho}_{i,j}$. It can be shown, by substituting eq. (4) in eq. (3), that the desired correlation consists a function of equivalent correlation, which can be expressed as:

$$\rho_{i,j} = f(\hat{\rho}_{i,j}, F_{X_i}, F_{X_j}) \quad (5)$$

Where F_{X_i} and F_{X_j} denote the specified marginal distributions. This relationship should be resolved for every pair $\hat{\rho}_{i,j}$ $(i \neq j)$ of the auxiliary PAR(1) process. The literature includes a variety of approaches to solve the latter equation, including Newton's method (Cario and Nelson, 1997, 1996; Li and Hammond, 1975), root-fnding methods (Chen, 2001), as well as, numerical integration and Monte-Carlo methods (Xiao, 2014). In this paper we employ a simple algorithm based on Monte-Carlo simulation and polynomial approximation proposed by Tsoukalas et al., (2017).

3. Case study

In order to illustrate the potential of SPARTA method we choose to employ a theoretical case study of an artificial univariate time series. Let us assume that we want to simulate an annual process $\{X_s\}$ consisted of 12 seasons. Furthermore, let us assume that each one has different marginal distribution and its parameters are *a priori* known. The specified distributions as well as their parameters are synopsized in Table 1. Furthermore, we assumed that the desired season-to-season correlation is equal to, $\boldsymbol{\rho} = [\rho_{12,1}, \rho_{1,2}, \dots, \rho_{t,t-1} \dots, \rho_{11,12}] =$ [0.7,0.6,0.3,0.5,0.6,0.7,0.5,0.6,0.7,0.8,0.7,0.6].

Since the marginal distributions and their parameters are already known the generation procedure reduces in to performing steps (b) – (e) of the procedure described in section 2. More specifically we employed PAR(1) as auxiliary model which is consisted of 12 parameters and hence, the double integral in eq. (4) had to be resolved 12 times. The performance assessment of SPARTA was based on its ability to capture the key statistical characteristics (i.e., mean, standard deviation, skewness, kurtosis and season-to-season correlation) of theoretical distributions as well as its ability to exactly reproduce the specified marginal distributions.

4. Results

To this end we employed SPARTA and simulated 5 000 years of the process $\{X_s\}$. As depicted in Figure 1, the model was able to accurately reproduce the seasonal mean and standard deviation with high precision where the two lines are almost indistinguishable. A similar behavior is observed when comparing the theoretical and simulated values of skewness and kurtosis. The latter behavior highlights the ability of the model to capture the key statistical characteristics of the understudy process even if different marginal models are established for each season.

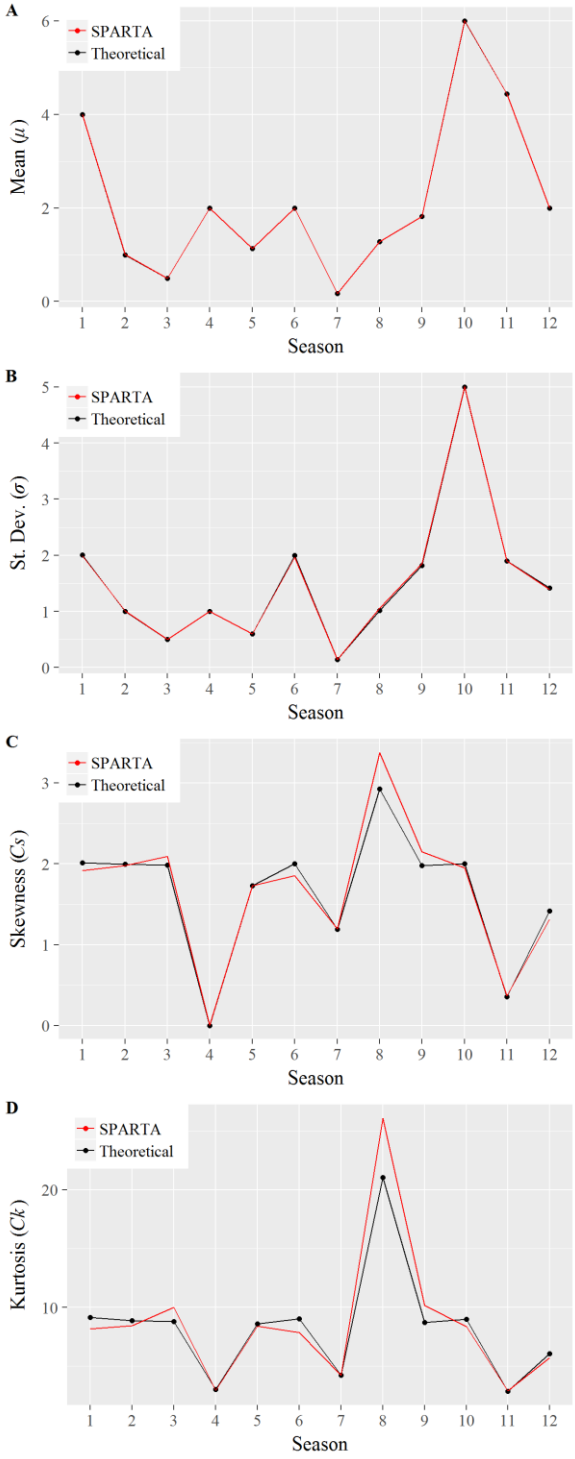

Figure 1: Comparison of theoretical and simulated values of seasonal **A**) mean (μ), **B**) standard deviation (σ), **C**) skewness (*Cs*) and **D**) kurtosis (*Ck*).

Table 1: Theoretical distributions and parameters of each season of the artificial time series 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	Beta	LoNo	Exp	PIII	Wei	Gam
	Theoretical Values											
a	1	1	1	2	0	1	1	0	0.55	1	2.5	2
b	2	-	2	1	0.5	2	5	0.7	-	1	5	1
c	2	-	-	-	-	-	-	-	-	5	-	-
	Simulated Values											
a	1.01	0.97	1.02	1.97	0.001	1.04	1.98	0.002	0.52	1.01	2.48	2.02
b	1.97	-	2.01	0.99	0.50	2.02	4.92	0.71	-	0.97	5.01	1.02
c	2.05	-	-	-	-	-	-	-	-	5.03	-	-

*Distribution abbreviations: PIII: Pearson III (*a* = shape, *b* = rate, *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); Beta: Beta (a = shape, b = shape).

Likewise, Figure 2 illustrates the performance of SPARTA in terms of reproducing the desired season-to-season correlation. Again, the theoretical and simulated values are almost undistinguishable. Furthermore, the identified equivalent correlation coefficients are depicted in the same graph in order to provide an insight to the reader.

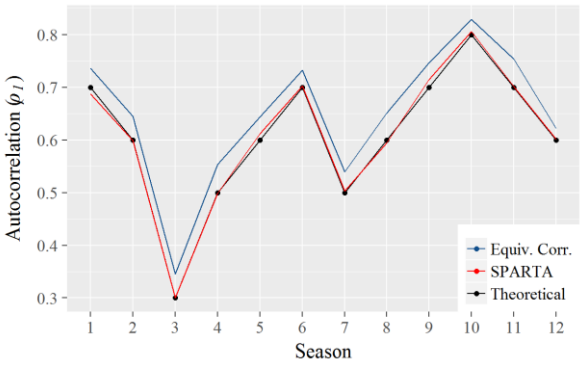

Figure 2: Comparison between theoretical (black line) and simulated (red line) season-to-season correlation ($\rho(1)$). The blue line illustrated the estimated equivalent correlation coefficients.

In order to further investigate the performance of the model we reverse-estimated the parameters of the distributions using the simulated data and the maximum likelihood method (MLE). Table 1 summarizes the estimated parameters which show a close agreement with their theoretical values. This can be also visually confirmed in Figure 3 where we compare the theoretical and simulated CDFs of two seasons (i.e., season 2 and 5). Again, the simulated data closely agree with the theoretical values, highlighting the "exactness" of the method in terms of reproducing the marginal distribution. Another notable characteristic of the model is that can by definition (through the use of eq. (1)) allows the avoidance of generating negative values. This is realized when the specified marginal distribution is positively bounded. For example, it is known that the exponential distribution (season 2) is bounded as follows, $x \in [0, \infty)$ therefore the lowest possible value that can be generated by the SPARTA method is zero. The same applies for the log-normal distribution which is defined for $x \in (0, \infty)$.

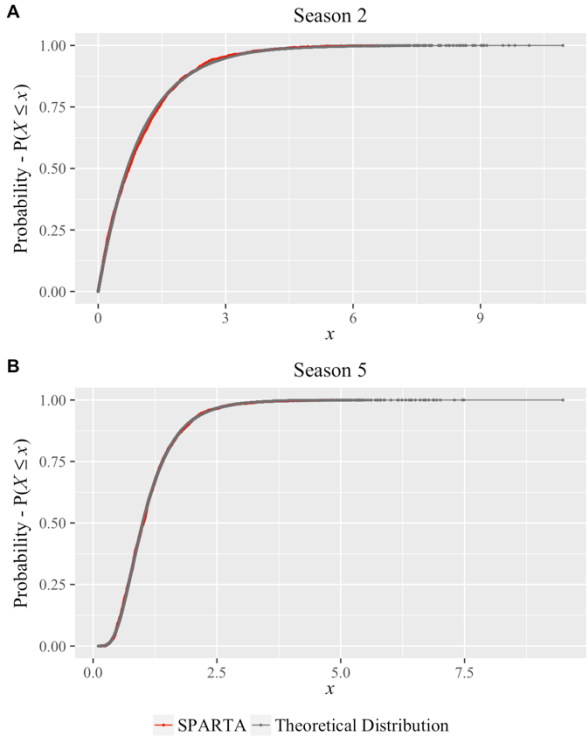

Figure 3: Comparison of theoretical and simulated cumulative density function (CDF) of **A**) season 2 and **B**) season 5 using the Weibull plotting position

5. Conclusions

In this work, we presented a novel cyclo-statonary model, termed Stochastic Periodic AutoRegressive To Anything (SPARTA) suitable for the simulation of periodic time series with arbitrary marginal distributions. The central idea of SPARTA lies into employing the Nataf's joint distribution model to capture the dependency among seasons and simultaneously exactly preserve their marginal distributions. The latter is attained with the use of an auxiliary periodic model from the PAR family with such parameters that after the mapping to the "real" domain attain the desired correlation structure. Apart from the obvious advantage of simulating data with exact marginal distribution, the proposed model, in contrast to the classic PAR models, can avoid the generation of negative values which have no physical meaning for hydrological time series. Another advantage of SPARTA is its parsimonious structure since it has the same number of parameters with a typical PAR model. The performance of SPARTA was assessed using a "toy" case study that involved the simulation of a periodic process exhibiting different marginal distribution for each season and seasonal correlation structure. SPARTA was able not only to reproduce the theoretical statistics and the temporal correlation structure but also reproduce the parameters of the prescribed marginal distributions. Finally, it can be argued that the flexibility of the proposed method, concerning the selection of different distributions and fitting methods, allows the incorporation of recent advances of statistical science within the domain of stochastic hydrology. Future work will be focused on extending SPARTA for multivariate simulation (Tsoukalas et al., 2017), as well as, coupling it with disaggregation techniques (e.g., Koutsoyiannis and Manetas, 1996).

6. References

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

Cario, M.C., Nelson, B.L., 1997. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Industrial Engineering 1–19.

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

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

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

Fiering, B., Jackson, B., 1971. Synthetic Streamflows, Water Resources Monograph. American Geophysical Union, Washington, D. C. doi:10.1029/WM001

Kottegoda, N.T., 1980. Stochastic water resources technology. Springer.

Koutsoyiannis, D., Manetas, A., 1996. Simple disaggregation by accurate adjusting procedures. Water Resources Research 32, 2105–2117. doi:10.1029/96WR00488

Koutsoyiannis, D., Yao, H., Georgakakos, A., 2008. Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 53, 142–164. doi:10.1623/hysj.53.1.142

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

Matalas, N.C., Wallis, J.R., 1976. Generation of synthetic flow sequences, Systems Approach to Water Management. McGraw-Hill, New York, New York.

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

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

Salas, J.D., Tabios, G.Q., Bartolini, P., 1985. Approaches to multivariate modeling of water resources time series. Journal of the American Water Resources Association 21, 683–708. doi:10.1111/j.1752-1688.1985.tb05383.x

Thomas, H.A., Fiering, M.B., 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. Design of water resource systems 459–493.

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

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