The mode of the climacogram estimator for a Gaussian Hurst-Kolmogorov process

Short Title: Same

Panayiotis Dimitriadis* and Demetris Koutsoyiannis

Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, 15880 Zographou, Greece

*corresponding author, email: pandim@itia.ntua.gr

Abstract

Geophysical processes are often characterized by long-term persistence. An important characteristic of such behaviour is the induced large statistical bias, i.e. the deviation of a statistical characteristic from its theoretical value. Here, we examine the most probable value (i.e. mode) of the estimator of variance to adjust the model for statistical bias. Particularly, we conduct an extensive Monte-Carlo analysis based on the climacogram (i.e. variance of the average process vs. scale) of the simple scaling (Gaussian Hurst-Kolmogorov) process, and we show that its classical estimator is highly skewed especially in large scales. We observe that the mode of the climacogram estimator can be well approximated by its lower quartile (25% quantile). To derive an easy-to-fit empirical expression for the mode we assume that the climacogram estimator follows a gamma distribution, an assumption strictly valid for Gaussian white noise processes. The results suggest that when a single timeseries is available it is advantageous to estimate the Hurst parameter using the mode estimator rather than the expected one. Finally, it is discussed that while the proposed model for mode bias works well for Gaussian processes, for higher accuracy and non-Gaussian processes one should perform a Monte-Carlo simulation following an explicit generation algorithm.

Keywords: statistical bias, long-term persistence, stochastic uncertainty, mode estimator, climacogram

1. Introduction

An important attribute characterizing geophysical processes is the high spatio-temporal dependence, in the sense that a random variable of such a process at a specific time or location strongly depends on several (even infinite) past, or of different location, random variables of the same process. This type of dependence requires long samples for its identification, which is a rare case in most natural processes, and thus, for the estimation of its parameters it is advised to use only up to the second-order statistics (Lombardo et al., 2014) and only in cases where very long samples are available to expand to higher orders. The above issues are further highlighted in Dimitriadis (2017), where several (overall thirteen) such processes with various lengths and physical properties expanding from small-scale turbulence to large-scale hydrometeorological processes are analyzed in terms of their long-term behaviour using massive databases and unbiased estimators of the second-order dependence structure. Interestingly, all the examined processes exhibited long-term-persistence, else known as Hurst-Kolmogorov (HK) behaviour (coined by Koutsoyiannis and Cohn, 2008), i.e. power-law decay of the autocorrelation function with lag (for a literature review on long-term persistent processes in hydrometeorology see also

O'Connell et al., 2016). Additionally, Koutsoyiannis (2011) provided a theoretical justification of the HK behaviour in geophysical processes showing that it is linked to the second-law of thermodynamics (i.e. entropy extremization), and specifically, the stronger the persistence of the dependence structure of a process, the higher the entropy of the process at large scales.

The identification of the dependence structure of a process can be highly affected by the sample uncertainty and statistical bias, where the true statistical properties (mean, variance etc.) of a statistic (e.g. variance) of a stochastic process may differ from the one estimated from a series with finite length. The deviations of the statistical characteristics from their true values should be taken into account not only for the marginal characteristics but also for the dependence structure of the process. Therefore, to correctly adjust the stochastic model to the observed series of the physical process we should account for the bias effect since all series are of finite (and often short) lengths.

The second-order properties can be similarly assessed by common stochastic tools such as the autocovariance function (a function of lag), power spectrum (a function of frequency), and variation of statistics (e.g. variance) of the averaged process vs. scale, a tool known as climacogram (Koutsoyiannis, 2010). It is shown that the latter estimator of the second-order dependence structure, as compared to the other two metrics, encompasses additional advantages in stochastic model building and interpretation from data; for example, it is characterized by smaller statistical uncertainty and easier to handle expressions of the statistical bias (Dimitriadis and Koutsoyiannis, 2015). Therefore, it is advisable that the sample uncertainty of the second-order dependence structure be tackled with the estimator with the lower variation, such as the climacogram. When multiple sample realizations (i.e. recorded series) are known, the handling of the statistical bias arising from a selected stochastic model may be based on the unbiased estimator of the expected value of the climacogram (Dimitriadis and Koutsoyiannis, 2018). However, when a single data series of observations is available for the model fitting (which is the case when geophysical processes are studied), it would be interesting to examine the mode of the climacogram, instead of the expected value; the two may differ in case of strong HK behaviour. This estimator is equivalent to a maximum-likelihood estimator (e.g. Kendziorski et al., 1999) for processes with zero (i.e. white noise) or short-term (e.g. Markov) dependence structure, while here we further extend it for HK processes (see also the work of Tyralis and Koutsoyiannis, 2011, for the expectation of the climacogram). It is noted that while the climacogram is often based on the second central moment (i.e. variance) other types of moments (e.g. raw, L-moments or K-moments; Koutsoyiannis, 2019) can be used to measure fluctuation in scale, and here, we focus on the central second-order climacogram (i.e. fluctuation measured by variance vs. scale).

2. Methods

In this section we present the applied methods, namely the climacogram estimator, the statistical bias expressions for the mode and expected values of the estimator and the algorithm for the stochastic synthesis of the Gaussian HK process for the Monte-Carlo analysis.

2.1. The climacogram

The analysis of a process through the variance of the averaged process vs. scale has been thoroughly applied in stochastic processes (e.g. Papoulis 1991; Vanmarcke, 2010). However, its importance to the analysis of the second-order dependence structure is highlighted mainly by more recent works (see a historical review in Koutsoyiannis, 2018). Also, the simple name *climacogram* allowed its further understanding through visualization; indeed, the term originates from the Greek *climax* (meaning scale) and *gramma* (meaning written; cf. the terms autocorrelogram for the autocorrelation, scaleogram for the power spectrum and wavelets).

It has been shown that the climacogram, treated as an estimator (rather than just a tool for the identification of long-term behaviour of the second-order dependence structure), has additional advantages from the more widely applied estimators of the autocovariance and power spectrum (Dimitriadis and Koutsoyiannis, 2015). Namely, the climacogram could provide a more direct, easy and accurate means to make diagnoses from data and build stochastic models in comparison to the power spectrum and autocovariance. For example, the climacogram, compared to other tools, has the lowest standardized estimation error for processes with short-term and long-term persistence, zero discretization error for averaged processes, simple and analytical expression for the statistical bias, always positive values, well defined and usually monotonic behaviour, smallest fluctuation of skewness on small scales while closest to zero skewness in larger scales, and mode closest to the expected (i.e. mean) value in large scales. Also, the climacogram is directly linked to the entropy production of a process (Koutsoyiannis, 2011; 2016). Furthermore, the climacogram expands the notion of *variance* by making it a function of *time scale* and is per se further expandable for statistics different from the central estimators of fluctuation (e.g. second raw moment, second L-moment vs. scale; Koutsoyiannis, 2019), for different characteristics of the estimator (e.g. mode, median), and even for moments of higher (e.g. third, fourth) orders (Dimitriadis and Koutsoyiannis, 2018). Recently, Koutsoyiannis (2019) extended the notion of climacogram for orders higher than two and showed how to substitute the joint moments of a process, allowing in this manner to tackle some limitations of the latter such as the discretization effect and statistical bias.

Symbolically, the climacogram is:

$$\gamma(k) \coloneqq \mathrm{var}[\underline{x}(k)] \tag{1}$$

where var[ ] denotes the variance and $\underline{x}(k) \coloneqq 1/k \int_0^k \underline{x}(t)\mathrm{d}t$ is the continuous-time process at scale k (in dimensions of time), which equals the discrete-one averaged in time intervals $\varDelta$, i.e. $\underline{x}_\kappa \coloneqq 1/\kappa \sum_{i=1}^{\kappa} \underline{x}_i$, in the discrete-time scale $\kappa = k/\varDelta$ (dimensionless natural number whereas for real numbers see adjustment in Koutsoyiannis, 2011).

2.2. The Gaussian long-term persistent process and its stochastic synthesis

The most common processes employed in geophysics, and particularly in hydrology, are the white noise process, the Markov process (with an exponential decay of the autocorrelation) and long-term persistent processes, which are characterized by a power-law decay of the climacogram (or equivalently of the autocorrelation) as a function of scale (or lag). A typical representative of the latter processes is the Gaussian HK process defined as:

$$\left(\underline{x}(k) - \mu\right) =_{\mathrm{d}} (k)^{(H-1)}\left(\underline{x}(1) - \mu\right) \tag{2}$$

where $=_{\mathrm{d}}$ denotes equality in distribution with μ the mean and $\gamma(k) = \gamma(\varDelta)/\kappa^{2-2H}$ the variance of the process for each scale k, H is the Hurst parameter ($0 < H < 1$) else defined as (Dimitriadis et al., 2016a) $H \coloneqq 1 + \frac{1}{2}\lim_{k\to\infty} \mathrm{d}\ln(\gamma(k)) / \mathrm{d}\ln k$; the quantity in the limit is the derivative of $\ln(\gamma(k))$ with respect to $\ln(k)$.

It is noted that this process has infinite variance at scale zero and thus, it should not be used to model the small scales of a physical process (in spite of the fact that the fractional-Gaussian-noise -fGn- process is widely used to model several processes at small scales; Koutsoyiannis et al., 2018). For the stochastic synthesis of the Gaussian HK model, we may use the sum of arbitrarily many independent Markov processes, thus expressing the target climacogram as (Dimitriadis and Koutsoyiannis, 2015):

$$\gamma(\kappa\Delta) = \sum_{i=1}^{l} \frac{2\lambda_i}{(\kappa\Delta/q_i)^2}\left(\kappa\Delta/q_i + e^{-\kappa\Delta/q_i} - 1\right) \tag{3}$$

where λ_i is the variance and q_i a time scale parameter for each Markov model i, and l the total number of Markov processes. Mandelbrot (1963) has shown that for $l \to \infty$ the above model can adequately describe an fGn (or else Gaussian HK) process for any generated length (see also Mandelbrot and Wallis, 1968; Mandelbrot and van Ness, 1968). Koutsoyiannis (2002) has analytically estimated the parameters of three AR(1) models (l = 3) to capture the HK process for $n < 10^4$. Dimitriadis and Koutsoyiannis (2015) have expanded this framework to the sum of arbitrarily many AR(1) models (abbreviated as SAR) for the generation of any type of process with autoregressive dependence structure and up to any number of scales, by using a suitable function with only two parameters, namely p_1 and p_2, that link the lag-1 autocorrelations of each Markov model, e.g. through the expression $q_i = p_1 {p_2}^{i-1}$, with i = 1, …,l and l often taken equal to the integer part of log(n)+1. For example, for $n = 10^6$ and H = 0.8, we have l = 7, p_1 =0.394 and p_2 =12.356 for a maximum standardized error between the true γ_t (Eqn. 2) and modelled γ_m (Eqn. 3) climacogram (i.e. $\max|(\gamma_t - \gamma_m)/\gamma_t|$ for all scales) equal to 0.009 (Table 1).

Table 1: Parameters p_1 and p_2 estimated to approximate different types of the N(0,1)-HK model (i.e. μ = 0 and $\gamma(\Delta)$ = 1) with l = 7 and $n \le 10^6$.

H	p_1	p_2	maximum error (standardized)
0.51	0.022	17.122	0.001
0.60	0.091	12.607	0.006
0.70	0.124	13.317	0.009
0.80	0.394	12.356	0.009
0.90	0.395	14.708	0.005
0.99	0.548	19.465	0.001

2.3. The mode of climacogram estimator and its statistical bias

The climacogram can be estimated from a sample through an estimator as similarly done for the estimators of the marginal moments. Here, for the climacogram we use a classical estimator:

$$\underline{\hat{\gamma}}(\kappa\Delta) = \frac{1}{\lfloor n/\kappa \rfloor - 1} \sum_{i=1}^{\lfloor n/\kappa \rfloor} \left(\underline{x}_i^{(\kappa)} - \underline{\bar{x}} \right)^2 \tag{4}$$

where $\lfloor n/\kappa \rfloor$ is the integer part of n/κ, $\underline{x}_i^{(\kappa)} = \sum_{l=\kappa(i-1)+1}^{\kappa i} \underline{x}_l / \kappa$ is the averaged process at scale $\kappa = k/\Delta$ for $i \in [1, \lfloor n/\kappa \rfloor]$, $\underline{\bar{x}} = \sum_{l=1}^{n} \underline{x}_l / n = \underline{x}_1^{(n)}$ is the sample average and n is the series length.

Since the above estimator is a random variable, it has a marginal distribution (see an illustration in Fig. 1). The true value of a statistical characteristic (e.g. variance) of a stochastic model may differ from the one estimated from a series with finite length. To correctly adjust the stochastic model to the observed series of the physical process one should account for the bias effect. An important question is how the statistical bias is generally handled through the second-order dependence structure in case of long-term persistent processes. Particularly, the selected stochastic model should be adjusted for bias before it is fitted to the sample dependence structure. It is noted that neglecting the bias effect in case of a long-term persistent process leads to underestimations of the stochastic model parameters such as the Hurst parameter, and to erroneous conclusions. Adjustment of the models for bias is usually done by equating the observed dependence structure to the expected value of the applied estimator. The alternative studied here is the mode, instead of the expected value, of the dependence structure, which represents the most probable value (and thus, the most expected) of the variance estimator at each scale.

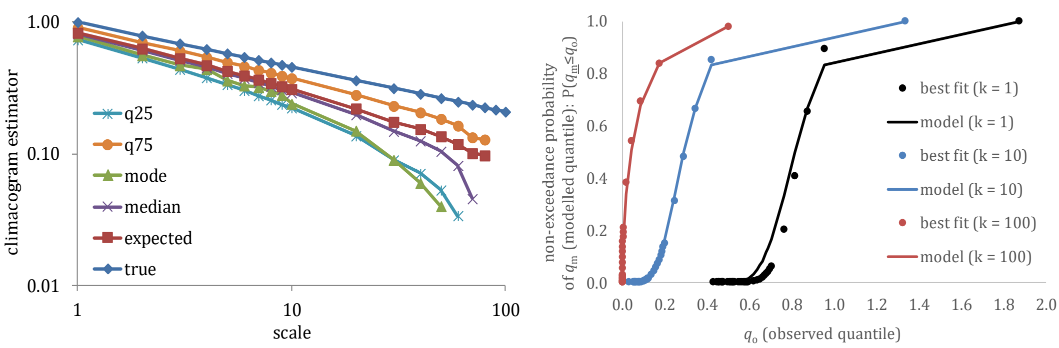

Figure 1: An illustration for a N(0,1)-HK (H = 0.83, n = 200) process of [left] how several statistical characteristics of the climacogram estimator vary with scale and [right] the observed quantile ($\underline{q}_o$) vs. the non-exceedance probability of the modelled quantile $P(\underline{q}_m \leq q_o)$, showing how the gamma distribution can adequately approximate the distribution of the climacogram estimator especially at large scales.

The statistical bias of an estimator is the difference of the expected value of an estimator from its true value (e.g. Papoulis, 1991). Thus, the bias of the climacogram is (e.g. Koutsoyiannis, 2011):

$$\mathrm{B_E}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] = \mathrm{E}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] - \gamma(\kappa\Delta) = \frac{(\kappa/n)\gamma(\kappa\Delta) - \gamma(n\Delta)}{1 - \kappa/n} \tag{5}$$

where $\mathrm{B_E}[\,]$ denotes the bias of the expected value of a statistical estimator of a process. Clearly, for the mean value of a process we have that $\mathrm{B_E}\left[\underline{\hat{\mu}}\right] = \mathrm{E}\left[\sum_{l=1}^{n} \underline{x}_l / n\right] - \mu = 0$.

Following the same rationale, we define an expansion of the notion of bias for the mode of the above estimator of the climacogram, i.e.:

$$\mathrm{B_M}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] = \mathrm{M}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] - \gamma(\kappa\Delta) \tag{6}$$

where $\mathrm{M}[\underline{x}] \coloneqq \arg\max[f(\underline{x})]$ denotes the mode of the variable $\underline{x}$ with density function $f(x)$. We refer to $\mathrm{B_M}[\,]$ as the mode bias.

For a Gaussian white noise process of length n and variance $\gamma(1)$ the distribution of its sample variance follows the gamma distribution $\Gamma\big((n-1)/2\,, 2\gamma(1)/(n-1)\big)$ (Cochran, 1934). The averaged process at scale κ, with sample length of n/κ and variance $\gamma(k) = \gamma(\Delta)/\kappa$, follows $\Gamma\big((n/\kappa-1)/2\,, 2\gamma(\Delta)/(n-\kappa)\big)$, with $\mathrm{M}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] = \gamma(\Delta)\frac{n-3\kappa}{\kappa(n-\kappa)}$ for $n/\kappa \geq 3$, else 0. Hence, for $n/\kappa \gg 3$ we have that $(n-3\kappa)/(n-\kappa) \approx 1$, and $\mathrm{M}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] \approx \mathrm{E}\left[\underline{\hat{\gamma}}(\kappa\Delta)\right] = \gamma(\Delta)/\kappa = \gamma(\kappa\Delta)$, i.e. zero bias. However, for long-term persistent processes the mode bias is non-zero and its analytical solution is no longer possible.

From the above results it becomes evident that the statistical bias always depends on the selected model and not on the data as commonly thought. For example, consider the Gaussian-HK process in the previous section with an autocorrelation function in discrete-time $\rho_\upsilon = 1/2\big(|\upsilon+1|^{2H} + |\upsilon-1|^{2H}\big) - |\upsilon|^{2H}$, where υ is the discrete-time lag. The bias of the autocorrelation is similarly defined as $\mathrm{B_E}\left[\underline{\hat{\rho}}(\upsilon)\right] = \mathrm{E}\left[\underline{\hat{\rho}}(\upsilon)\right] - \rho(\upsilon)$, and thus, depends on the model parameter H. It is noted that the above apply even to the so-called non-parametric models, since they also involve estimation from data, and thus, these models should be similarly adjusted for statistical bias to avoid underestimation of the process variability during a Monte-Carlo simulation.

For simplicity, and without loss of generality, we set Δ = 1 for the rest of the analysis. It is evident that $\mathrm{B_M}\left[\underline{\hat{\gamma}}(\kappa)\right] \leq \mathrm{B_E}\left[\underline{\hat{\gamma}}(\kappa)\right] \leq 0$ or else $\left|\mathrm{B_E}\left[\underline{\hat{\gamma}}(\kappa)\right]\right| \leq \left|\mathrm{B_M}\left[\underline{\hat{\gamma}}(\kappa)\right]\right|$, since the sample variance is positively skewed, i.e. $\mathrm{E}\left[\underline{\hat{\gamma}}(\kappa)\right] \geq \mathrm{M}\left[\underline{\hat{\gamma}}(\kappa)\right]$, and the equality holds when $n \rightarrow \infty$, where the variance of the sample variance is zero for an ergodic process. A preliminary analysis of common HK-type processes has shown that the mode climacogram is close to the low quartile (25% quantile) of the marginal distribution of variance at each scale (Dimitriadis et al., 2016c; Gournary, 2017). Therefore, when the mode of the variance estimator is of interest, we may use a Monte-Carlo technique (as described in the next section) to accurately estimate the mode bias or, in case the marginal distribution of the climacogram is known, to calculate the 25% quantile at each scale to approximate the mode bias.

3. Monte-Carlo analysis for the mode of the variance estimator

We perform Monte-Carlo experiments over the N(0,1)-HK model for a wide range of Hurst parameters H (i.e. 0.5 to 0.95), and for a wide range of series lengths n (i.e. 20 to 2000). Specifically, we produce a number (N) of synthetic series through the SAR model described in section 2.2, where N depends on the sample mean value to reach the expected one at scale κ = $n/10$ based on the rule of thumb when using the climacogram as shown in Dimitriadis and Koutsoyiannis (2015). We found that for $N \approx 10^6/n^{2-2H}$ the standardized error between the theoretical expected value and the sample one (Eqn. 5) is lower than 1% at scale κ = $n/10$. In this way, the mode is expected also to be well preserved with a similar error. However, caution should be given to the selection of the sample mode estimator to ensure that its variance permits a robust estimation of the true value of the mode. Since the distribution function of the estimator of variance is unknown for long-term persistent processes, and given that the mode value is the most probable to occur within the sample, we calculate the sample mode from each simulated series by finding the most probable value with an accuracy of two decimal digits. Specifically, we round up each value of the time series, and for each scale, to the second decimal digit, and we estimate the most probable value of the rounded time series (for higher accuracies a larger N was required). Also, other estimators for the sample mode (e.g. Bickel and Fruwirth, 2006) could be used and compared to the proposed one in future research to optimize the performance of the analysis.

Here, to derive an easy-to-fit empirical expression to approximate the mode bias, we adopt the assumption that the above distribution is nearly gamma for smaller scales (see also a similar analysis in Gournary, 2017, and Dimitriadis et al., 2018). Using the results from the Monte-Carlo analysis we then evaluate the parameter of the gamma distribution for each H, n and κ, and we build a model for the mode, which we later test its performance. Although the true autocorrelation function of the averaged process for a long-term persistent process does not vary with scale, the sample autocorrelation will be also prone to bias (e.g. Dimitriadis and Koutsoyiannis, 2015) affecting the distribution function of the sample variance at each scale. To minimize the sample error for the fitting of the two-parameter gamma distribution we use the theoretical expression for the expected value of the sample climacogram, i.e. $\mathrm{E}[\underline{\hat{\gamma}}(\kappa)]$, and the variance of the sample climacogram, i.e. $\mathrm{Var}[\underline{\hat{\gamma}}(\kappa)]$, as evaluated from the Monte-Carlo analysis, which exhibits the lowest variability in estimation among the four central moments (Dimitriadis and Koutsoyiannis, 2018, Fig. 2). Based on these two measures, we estimate the two parameters of the gamma distribution.

We first set the scale parameter of the gamma distribution such as to simulate the sample ratio of the aforementioned parameters, i.e. $b(H,n,\kappa) = \mathrm{Var}[\underline{\hat{\gamma}}(\kappa)]/\mathrm{E}[\underline{\hat{\gamma}}(\kappa)]$ and so, the shape parameter can be also estimated as $a(H,n,\kappa) = \mathrm{E}[\underline{\hat{\gamma}}(\kappa)]/b(H,n,\kappa)$.

We observe (e.g. Fig. 2) that for $a(H,n,\kappa) > 1$ the shape parameter $a(H,n,\kappa)$ is approximately proportional, by a function $c(H)$, to the corresponding shape parameter for the white noise process $a(0.5,n,\kappa) = (n/\kappa - 1)/2$ raised to a function $p(H)$, i.e.:

$$a(H,n,\kappa) = c(H)\big((n/\kappa - 1)/2\big)^{p(H)} \quad (7)$$

where $a(H,n,\kappa) > 1$ is a function corresponding to the shape parameter of the gamma distribution function, while for $a(H,n,\kappa) \leq 1$ or $\kappa \gtrsim n/3$, the mode is considered close to zero.

The two functions of the above expression are fitted as (Fig. 2):

$$c(H) = 1.68(H - 0.5)^2 - 0.3025(H - 0.5) + 1 \quad (8)$$

and

$$p(H) = -2.4865(H - 0.5)^2 + 0.1485(H - 0.5) + 1 \quad (9)$$

The above two adjustments allow to empirically express the mode of the climacogram estimator as a function of H, n and κ:

$$\mathrm{M}\left[\hat{\underline{\gamma}}(\kappa)\right] = (a(H,n,\kappa) - 1)b(H,n,\kappa) = \left(1 - 1/a(H,n,\kappa)\right)\mathrm{E}\left[\hat{\underline{\gamma}}(\kappa)\right] \tag{10}$$

It is noted that based on the above assumptions the standard deviation, and the skewness and excess kurtosis coefficients of the climacogram estimator can be estimated as $b(H,n,\kappa)\sqrt{a(H,n,\kappa)}$, $2/\sqrt{a(H,n,\kappa)}$, and $6/a(H,n,\kappa)$, respectively. Since $a(H,n,\kappa) \leq a(0.5,n,\kappa)$ all the above measures will be larger than those in case of a white noise process.

The above expression can approximate the mode by an absolute difference of 0.005 from the Monte-Carlo estimates, while for better approximations it is advised to implement a new Monte-Carlo analysis (see also discussion and application in sect. 4). Interestingly, the standardized error between the mode and expected values of the estimator, i.e. $\varepsilon = \left|\mathrm{E}[\hat{\underline{\gamma}}(\kappa)] - \mathrm{M}[\hat{\underline{\gamma}}(\kappa)]\right| / \mathrm{E}[\hat{\underline{\gamma}}(\kappa)]$, is calculated from the Monte-Carlo analysis to reach a maximum value of 67% corresponding to cases with $H \geq 0.6$ and $n/\kappa \leq 10$, while for the white noise process it can be theoretically estimated as $\varepsilon = 2/(n/\kappa - 1)$, which for $\kappa = n/10$ is approximately 20%.

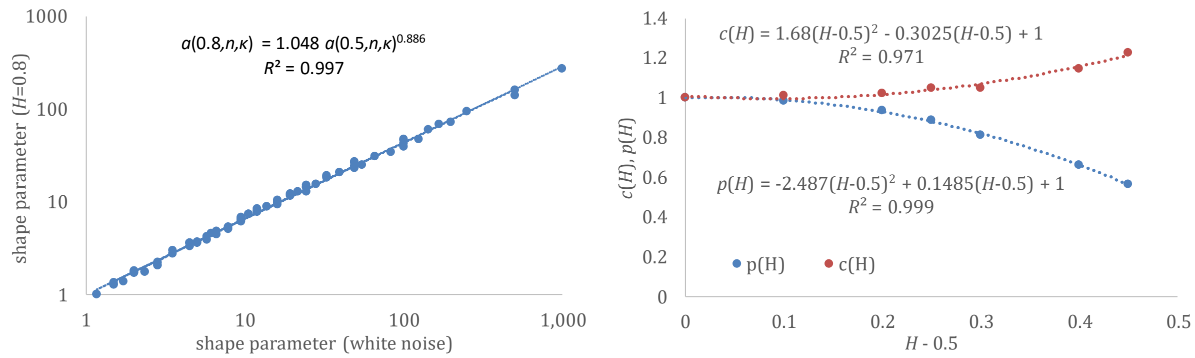

Figure 2: [left] The shape parameter assuming a gamma distribution for the mode estimator of the climacogram of an N(0,1)-HK process (for H = 0.8 and for all n and κ simulated in the Monte-Carlo analysis) versus the theoretical shape parameter of the white noise process. [right] Proposed model for the $c(H)$ and $p(H)$ functions for all examined H from the Monte-Carlo analysis.

4. Applications to annual streamflow

For illustrations of possible implications of the above results, we apply a stochastic analysis based on the expected and the mode values of the climacogram to a streamflow process at the Peneios river (Thessaly, Greece), where a historical streamflow annual time series is available at the upstream station of Ali Efenti with only a 13 years length (for more information on the study area see Dimitriadis et al., 2016b). For the identification of the stochastic model we adjust for statistical bias and, in particular, we fit the mode of the estimator rather than its expectation. It is noted that the proposed empirical model for the mode bias (Eqn. 10) is derived from a Monte-Carlo analysis for sample lengths of $n \geq 20$, and so for this application we perform a new Monte-Carlo analysis to fit the observed climacogram for scales $1 \leq \kappa \leq n/10$ (rule of thumb; Dimitriadis and Koutsoyiannis, 2015) and so here, for the first two scales (Fig. 3). We find that an HK model can adequately simulate the observed standardized climacogram, i.e. $\hat{\gamma}(\kappa)/\hat{\gamma}(1)$, with H = 0.9. We also estimate the Hurst parameter with the expectation of the estimator, and we find $H' \approx 0.8$ and $H'' \approx 0.7$, with or without adjusting for bias, respectively. Evidently, both latter values underestimate the long-term persistence behaviour (Fig. 3).

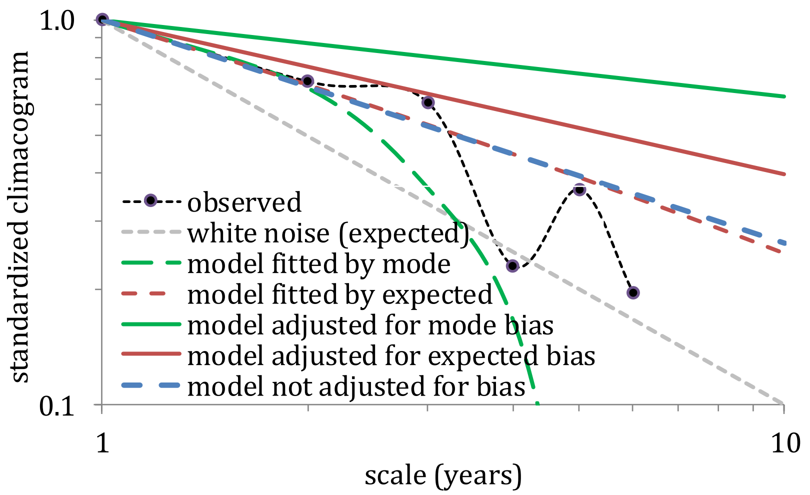

Figure 3: Standardized climacogram estimations of the observed standardized time series (black line), the white noise model (grey line), and the three fitted N(0,1)-HK stochastic processes: (a) adjusting for bias of the mode of the estimator (green line), i.e. $\mathrm{M}[\underline{\hat{\gamma}}(\kappa)]/\mathrm{M}[\underline{\hat{\gamma}}(1)]$, , and of its expectation (red line), i.e. $\mathrm{E}[\underline{\hat{\gamma}}(\kappa)]/\mathrm{E}[\underline{\hat{\gamma}}(1)]$, and (b) not adjusting for bias (blue line), i.e. $\underline{\hat{\gamma}}(\kappa) = \gamma(\kappa)$, also corresponding to the non-parametric model configuration.

It is noted that the dependence structure of a process (e.g. streamflow) will have a small effect at the risk imposed by the expected number of peaks over threshold (e.g. for the design of a dam or for flood risk mapping) as compared to the effect of the marginal distribution of the process (Volpi et al., 2015; Serinaldi and Kilsby, 2018). However, the dependence structure will have a great effect (especially for processes with long-term behaviour) at the duration of successive peaks over threshold (e.g. maximum duration of wet/dry periods or of flood inundation), which may highly affect urban as well as agricultural areas and insurance policies (e.g. Serinaldi and Kilsby, 2016; Goulianou et al., 2019). To illustrate this, we generate an adequate number N (see sect. 3) of HK synthetic timeseries with H = 0.5 ($N = 5\times10^3$), H = 0.7 ($N = 4\times10^4$), H = 0.8 ($N = 10^5$) and H = 0.9 ($N = 3\times10^5$). For convenience and simplification, we assume a N(0,1) distribution for all processes. We then estimate the expected frequency of the number of peaks over various thresholds (PoT) as well as the expected frequency of the maximum duration of successive peaks over various thresholds (MdT), and we standardize them with the PoT and MdT values of the white noise process (Fig. 4). We find that the MdT varies with threshold and long-term persistence while the PoT stays almost unaffected by both. Additional analyses and quantifications on the reflection of long-term term persistence in terms of clustering in time can be found in Iliopoulou and Koutsoyiannis (2019).

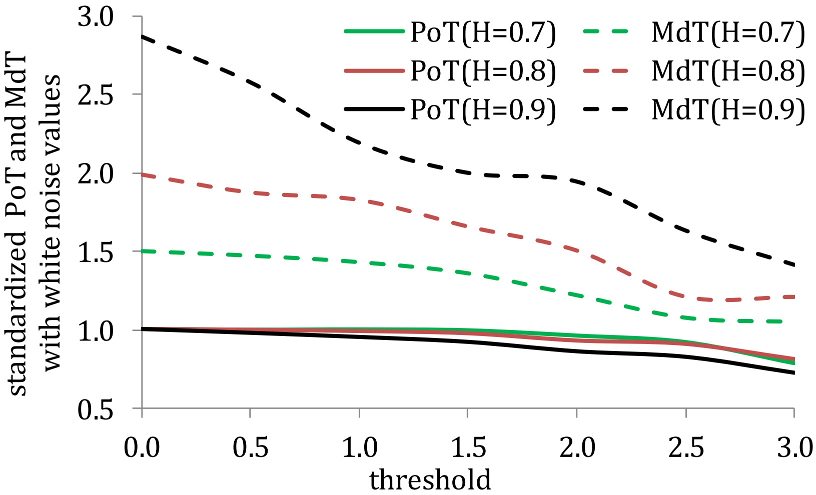

Figure 4: Expected frequency of peak over threshold (PoT) and expected maximum duration of successive peaks over threshold (MdT) standardized with the PoT and MdT values of the N(0,1) white noise process for various HK-N(0,1) processes.

The results from this study suggest that the sample estimator of the variance can be skewed even for long samples in the presence of long-term persistence behaviour as opposed to the white noise process. Therefore, the mode is different from the expectation and more suitable to use in estimation. We propose that, when a single recorded series is available and a Gaussian HK process is fitted with small sample size and relatively high Hurst parameter, it is advantageous to employ the mode of the estimator as calculated from the empirical model of Eqn. 10, rather than its expectation (Eqn. 5), so as to avoid underestimation of the Hurst parameter (and thus, the uncertainty of the process). In case of a non-Gaussian distribution, larger accuracy or a different estimator of the second-order dependence structure (e.g. other climacogram estimator, autocovariance, power spectrum, variogram etc.), we should employ the Monte-Carlo technique and test whether the mode of the estimator used is close enough to its expected value. If this is true then the expected value can be used to adjust the model for bias, whereas if the two values vary then for the model should be adjusted for bias based on the mode estimator. For Monte-Carlo analysis of a non-Gaussian correlated process an explicit algorithm should be preferred (Dimitriadis and Koutsoyiannis, 2018) since the mode value is expected to highly depend on higher-order moments in case of long-term persistent processes.

5. Conclusions and discussion

Awareness of uncertainty in assessing the dependence structure of a process is of paramount importance as it may critically affect the interpretation of results. Estimation uncertainty may introduce large statistical bias, which can be additionally magnified in the presence of long-term persistence (Dimitriadis and Koutsoyiannis, 2015). In addition, if the uncertainty is underestimated then a regular cluster of events could be erroneously regarded as an extreme cluster. Although the mode of the examined classical estimator for variance is close to its expectation for small Hurst parameters and large lengths, we show that for larger values of the Hurst parameter and small sample lengths, equating the expected climacogram to the observed one may lead to underestimation of the long-term persistence and thus, the uncertainty of the process.

We propose that when the available series have short lengths or when the empirical Hurst parameter is estimated larger than 0.5, we should always account for statistical bias. Particularly for the bias adaptation, when information is available on only a single series/realization of the process, it is advantageous to equate the mode instead of the expectation of the climacogram estimator to the sample values. Interestingly, in case of a N(0,1)-HK process, the absolute difference between the mode and expected values of the estimator is calculated (from a Monte-Carlo analysis performed in this study) to reach a maximum value of 67% of the expected value, corresponding to cases with $H \geq 0.6$ and $n/\kappa \leq 10$, while for the white noise process is approximately 20% for $\kappa = n/10$. In cases of different stochastic processes or estimators or when a larger accuracy of the mode bias is of interest, one should employ a Monte-Carlo technique through an explicit generation algorithm (Dimitriadis and Koutsoyiannis, 2018) to estimate the mode climacogram estimator or use the lower quartile (25% quantile) of the estimator (in case its distribution is known) as an approximation.

From the Monte-Carlo analysis performed in this study, it is also observed that for a N(0,1)-HK process with variance $\gamma(\kappa) = \gamma(1)/\kappa^{2-2H}$, and for large n and small n/κ, the distribution of the climacogram estimator tends to that of $\Gamma\big((n/\kappa - 1)/2\,, 2\gamma(\kappa)/(n/\kappa - 1)\big)$, with a mean value of $\gamma(\kappa)$, i.e. zero bias. However, given the estimation uncertainty present in records exhibiting persistence, the autocorrelation of the averaged process is independent of the scale, and thus, the above distribution will never be truly reached. The underestimation of the persistence of the parent process has also critical implications for the estimation of the properties of its extremes, as it was shown that the maximum duration of successive peaks over threshold is greatly affected by the degree of dependence. Additional analyses and quantifications on the reflection of long-term term persistence in terms of clustering in time can be found in Iliopoulou and Koutsoyiannis (2019).

A final remark for discussion, considering the etymology of the terms, is that the expected value of a random process is less expected to occur than its mode (i.e. most probable value; a term coined by Pearson, 1895, p. 345), where only in symmetrical distributions the two coincide. Therefore, when only one value is known (here, only one realization of the climacogram estimator), it is more accurate to fit the model and evaluate the Hurst parameter based on the proposed mode estimator rather than the expected one.

Acknowledgment

The Authors would like to thank the Editor Luigi Berardi for handling the paper, one anonymous Reviewer for useful comments, and Federico Lombardo for his fruitful discussion, comments and suggestions that helped us improve the paper.

Code availability

The MATLAB script for the SAR generation algorithm is available as well as the script for a fast estimation algorithm of the sample climacogram in very long timeseries and in many scales.

References

Bickel, D.R. and Fruwirth, R., On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications, *Computational Statistics & Data Analysis*, 50: 3500–3530, 2006.

Cochran, W. G., The distribution of quadratic forms in a normal system, with applications to the analysis of covariance, *Mathematical Proceedings of the Cambridge Philosophical Society*, 30 (2): 178–191. doi:10.1017/S0305004100016595, 1934.

Dimitriadis, P., Hurst-Kolmogorov dynamics in hydrometeorological processes and in the microscale of turbulence, PhD thesis, 167 pages, *National Technical University of Athens*, 2017.

Dimitriadis, P., and D. Koutsoyiannis, Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes, *Stochastic Environmental Research & Risk Assessment*, 29 (6), 1649–1669, 2015.

Dimitriadis, P., and D. Koutsoyiannis, Stochastic synthesis approximating any process dependence and distribution, *Stochastic Environmental Research & Risk Assessment*, 32 (6), 1493–1515, doi:10.1007/s00477-018-1540-2, 2018.

Dimitriadis, P., D. Koutsoyiannis, and P. Papanicolaou, Stochastic similarities between the microscale of turbulence and hydrometeorological processes, Hydrological Sciences Journal, 61 (9), 1623–1640, doi:10.1080/02626667.2015.1085988, 2016a.

Dimitriadis, P., A. Tegos, A. Oikonomou, V. Pagana, A. Koukouvinos, N. Mamassis, D. Koutsoyiannis, and A. Efstratiadis, Comparative evaluation of 1D and quasi-2D hydraulic models based on benchmark and real-world applications for uncertainty assessment in flood mapping, *Journal of Hydrology*, 534, 478–492, 2016b.

Dimitriadis, P., N. Gournari, and D. Koutsoyiannis, Markov vs. Hurst-Kolmogorov behaviour identification in hydroclimatic processes, *European Geosciences Union General Assembly*, Vol. 18, EGU2016-14577-4, 2016c.

Dimitriadis, P., N. Gournary, A. Petsiou and D. Koutsoyiannis, How to adjust the fGn stochastic model for statistical bias when handling a single time series; application to annual flood inundation, *13th Hydroinformatics Conference*, 1-6 July 2018, Palermo, Italy, 2018.

Goulianou, T., K. Papoulakos, T. Iliopoulou, P. Dimitriadis, and D. Koutsoyiannis, Stochastic characteristics of flood impacts for agricultural insurance practices, *European Geosciences Union General Assembly*, Vol. 21, EGU2019-5891, 2019.

Gournary, N., *Probability distribution of the climacogram using Monte Carlo techniques*, Diploma thesis, 108 pages, Department of Water Resources and Environmental Engineering – National Technical University of Athens, Athens, July 2017 (in Greek).

Iliopoulou, T., and D. Koutsoyiannis, Revealing hidden persistence in maximum rainfall records, *Hydrological Sciences Journal*, doi.org/10.1080/02626667.2019.1657578, 2019.

Koutsoyiannis, D., The Hurst phenomenon and fractional Gaussian noise made easy, Hydrological Sciences Journal, 47 (4), 573–595, 2002.

Koutsoyiannis, D., HESS opinions "A random walk on water", *Hydrology and Earth System Sciences*, 14, 585–601, 2010.

Koutsoyiannis, D., Hurst-Kolmogorov dynamics as a result of extremal entropy production, *Physica A: Statistical Mechanics and its Applications*, 390 (8), 1424–1432, 2011.

Koutsoyiannis, D., Generic and parsimonious stochastic modelling for hydrology and beyond, *Hydrological Sciences Journal*, 61 (2), 225–244, 2016.

Koutsoyiannis, D., Climate change impacts on hydrological science: A comment on the relationship of the climacogram with Allan variance and variogram, *ResearchGate*, 2018.

Koutsoyiannis, D., Knowable moments for high-order stochastic characterization and modelling of hydrological processes, *Hydrological Sciences Journal*, 2019.

Koutsoyiannis, D., and T.A. Cohn, The Hurst phenomenon and climate (solicited), *European Geosciences Union General Assembly*, Vol. 10, Vienna, 11804, doi:10.13140/RG.2.2.13303.01447, European Geosciences Union, 2008.

Koutsoyiannis, D., Dimitriadis, P., Lombardo, F, and Stevens, S., *From fractals to stochastics: Seeking theoretical consistency in analysis of geophysical data*, Advances in Nonlinear Geosciences, edited by A.A. Tsonis, 237–278, Springer, 2018.

Kendziorski, C.M., J.B. Bassingthwaighte, and P.J. Tonellato, Evaluating maximum likelihood estimation methods to determine the Hurst coefficient, *Physica A*, V. 273, 3-4, pp. 439–451, 1999.

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

Mandelbrot, B.B., The Variation of Certain Speculative Prices, *J. Bus.*, 36, 394–419, 1963.

Mandelbrot, B.B. and Wallis, J.R., Noah, Joseph and operational hydrology, *Water Resour. Res.*, 4, 909–918, 1968.

Mandelbrot, B.B., and Van Ness, J.W., Fractional Brownian Motions, Fractional Noises and Applications, *SIAM Rev.*, 10, 422–437, 1968.

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

Papoulis, A., Probability, *Random Variables and Stochastic Processes*, 3rd edn., McGraw-Hill, New York, 1991.

Pearson, K., Contributions to the mathematical theory of evolution—II, Skew variation in homogeneous material, Philosophical Transactions of the Royal Society of London, 186, 343-414, 1895 (available at https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1895.0010).

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

Serinaldi, F., and Kilsby, C.G., Unsurprising Surprises: The Frequency of Record-breaking and Over-threshold Hydrological Extremes Under Spatial and Temporal Dependence, *Water Resources Research*, 54(9), 6460-6487, 2018.

Tyralis, H., and D. Koutsoyiannis, Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process, *Stochastic Environmental Research & Risk Assessment*, 25 (1), 21–33, 2011.

Vanmarcke, E., *Random Fields: Analysis and Synthesis*, World Scientific, New Jersey, USA, 2010.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D., One hundred years of return period: Strengths and limitations, *Water Resources Research*, 51(10), 8570-8585, 2015.