Climacogram vs. autocovariance and power spectrum in stochastic modelling for Markovian and Hurst-Kolmogorov processes

Panayiotis Dimitriadis* and Demetris Koutsoyiannis

Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, 158 80 Zographou, Greece

*corresponding author, email: pandim@itia.ntua.gr, tel.: +30-210-772-28-38, fax: +30-210-772-28-32.

Abstract

Three common stochastic tools, the climacogram i.e. variance of the time averaged process over averaging time scale, the autocovariance function and the power spectrum are compared to each other to assess each one's advantages and disadvantages in stochastic modelling and statistical inference. Although in theory all three are equivalent to each other (transformations one another expressing second order stochastic properties), in practical application their ability to characterize a geophysical process and their utility as statistical estimators may vary. In the analysis both Markovian and non Markovian stochastic processes, which have exponential and power-type autocovariances, respectively, are used. It is shown that, due to high bias in autocovariance estimation, as well as effects of process discretization and finite sample size, the power spectrum is also prone to bias and discretization errors as well as high uncertainty, which may misrepresent the process behaviour (e.g. Hurst phenomenon) if not taken into account. Moreover, it is shown that the classical climacogram estimator has small error as well as an expected value always positive, well-behaved and close to its mode (most probable value), all of which are important advantages in stochastic model building. In contrast, the power spectrum and the autocovariance do not have some of these properties. Therefore, when building a stochastic model, it seems beneficial to start from the climacogram, rather than the power spectrum or the autocovariance. The results are illustrated by a real world application based on the analysis of a long time series of high-frequency turbulent flow measurements.

Keywords: stochastic modelling; climacogram; autocovariance; power spectrum; uncertainty; bias; turbulence

1. Introduction

The power spectrum (or else spectral density) was introduced as a tool to estimate the distribution of the power (i.e. energy over time) of a sample over frequency, more than a century ago by Schuster (Stoica and Moses, 2004, p. xiii). Since then, various methods have been proposed and used to estimate the power spectrum, via the Fourier transform of the time series (periodogram) or its autocovariance or autocorrelation functions (for more information on these methods see in Stoica and Moses, 2004, ch. 2 and Gilgen et al., 2006, ch. 9). Most common (and also used in this paper) is that of the autocovariance which corresponds to the definition of the power spectrum of a stochastic process (for details, see sect. 2.3). However, this accurate mathematical definition lacks immediate physical interpretation since the Fourier transform of a function is nothing more than a mathematical tool to represent the function in the frequency domain in order to identify any periodic patterns which are not easily tracked in the time domain.

Several researchers have tried in the past to evaluate the statistical estimator of the power spectrum concluding that its major disadvantage is that of its large variance (Stoica and Moses, 2004, p. xiv). Notably, this variance is not reduced with increased sample size (Papoulis, 1991, p. 447). To remedy this, several mathematical smoothing techniques (e.g. windowing, regression analysis, see Stoica and Moses, 2004, ch. 2.6) have been developed. In cases of short datasets, trend-line approaches are most commonly used to obtain a very rough estimation of the model behaviour or simple rules to distinguish exponential and power-type behaviours (e.g., Fleming, 2008). In cases of long datasets, the most commonly used approach is the windowing (data partitioning), also known as the Welch approach, where a certain window function (the simplest of which is the Bartlett window) is applied to nearly independent segments. In the latter method, one has first to divide the sample into several segments (but only after insuring these segments have very small correlations between them), to calculate the power spectrum for each segment and then to estimate the average. Assuming that the process is stationary, this average will be the power spectrum estimate. Unfortunately, the more segments we divide the sample into, the more the cross-correlations between segments are increasing as well as the more we lose in low frequency values (since the lowest frequency is determined by the length of the segments). Thus, this method could be indeed a robust one, but only for a very long sample (which is a rare case in geophysics), only when there is no interest in the low frequency values (which can reveal large-scale behaviours) and only for an unbiased power spectrum estimator or at least for an 'a priori' known bias, e.g. via an analytical equation (which, as we will show in this study, is rarely the case). Based on these limitations, Dimitriadis et al. (2012) and Koutsoyiannis (2013a, b) provided examples where this smoothing technique fails to detect the large scale behaviour (i.e. Hurst phenomenon), gives small scale trends that are completely different from the ones characterizing the stochastic model and have several numerical calculation problems that could cause misinterpretation (see sect. 4 and Fig. 10d for an illustrative example of the limitations of this method). These all are due to the fact that the power spectrum estimator is biased and it is difficult to estimate this bias analytically. Nevertheless, the power spectrum is a useful tool to analyze a sample in harmonic functions and so, to detect any dominant frequencies (this is the reason behind harmonic analysis introduced by J. Fourier, 1822).

In this paper, we investigate the bias in power spectrum estimator (evaluated via the autocovariance) which are caused by the bias of autocovariance, the finite sample size and discretization of the continuous-time process, complementing earlier studies (e.g. Stoica and Moses, 2004, ch. 2.4). We also examine the asymptotic behaviour when the sample size tends to infinity, investigating the question whether or not the discrete power spectrum estimator is asymptotically unbiased or not. We perform similar investigations for the climacogram, a term coined by Koutsoyiannis (2010) to describe the variance of the time averaged process as a function of time scale. The concepts of autocovariance, power spectrum and climacogram are examined using both exponential and power-type autocovariance, as well as combinations thereof, in order to obtain representative results for most types of geophysical processes.

In sect. 2, we give the definitions of the concepts used in the paper and in sect. 3, we investigate the estimation of the climacogram, the autocovariance and the power spectrum for some characteristic processes, and we compare their classical estimators based on illustrative examples. In sect. 4, we present an application of these stochastic tools to a small scale turbulent process and propose certain practices to be used in stochastic modelling. Finally, in sect. 5 we summarize the analyses and derive some conclusions.

2. Definitions and notations

Stochastic processes are families of random variables (denoted as $\underline{x}(t)$, where underlined symbols denote random variables and t denotes time) that are often used to represent the temporal evolution of natural processes. Natural processes as well as their mathematical representation as stochastic processes evolve in continuous time. However, observed time series from these processes are characterized by a sampling time interval D, often fixed by the observer and a response time Δ of the instrument (Fig. 1). The time constants D and Δ affect the estimation of the statistical properties of the continuous time process. Two special cases, $\Delta \to 0$ and $D = \Delta$, are analyzed by Koutsoyiannis (2013a) who shows that in most tasks the differences are small and thus, here we will focus only on the case $D = \Delta > 0$ that is also practical for samples with small D (the Markovian process for any D and Δ, in terms of its autocovariance, is shown in sect. 4 of the supplementary material, abbreviated as SM). Thus, the discrete time stochastic process $\underline{x}_i^{(\Delta)}$, for $D = \Delta > 0$, can be calculated from $\underline{x}(t)$ as:

$$\underline{x}_i^{(\Delta)} = \frac{\int_{(i-1)\Delta}^{i} \underline{x}(\xi)\mathrm{d}\xi}{\Delta} \tag{1}$$

where $i \in [1, n]$ is an index representing discrete time, $n = \lfloor T/\Delta \rfloor$ is the total number of observations and $T \in [0, \infty)$ is time length of observations.

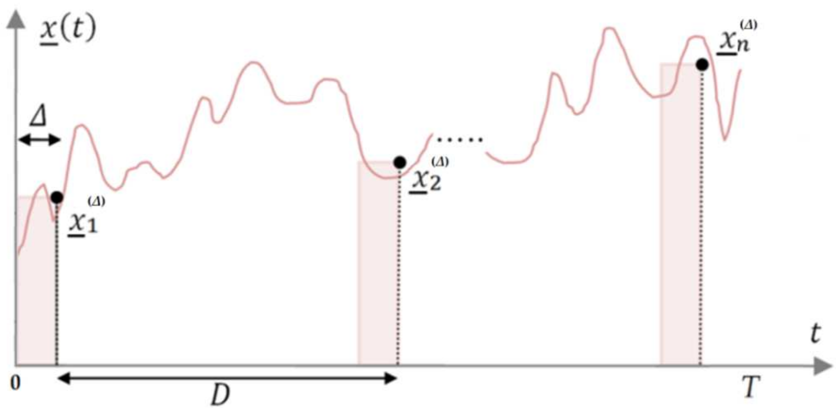

Figure 1: An example of a continuous time process sampled at time intervals D for a total period T and with instrument response time Δ.

2.1 Climacogram

The climacogram (Koutsoyiannis, 2013a) comes from the Greek word climax (meaning scale). It is defined as the (plot of) variance of the averaged process $\underline{x}(t)$ (assuming stationary) versus averaging time scale m and is symbolized by $\gamma(m)$. The climacogram is useful for detecting the long term change (or else dependence, persistence, clustering) of a process. This can be quantified through the Hurst coefficient H, which equals the half of the slope of the climacogram in a log-log plot, as scale tends to infinity, plus 1. For sufficiently large scales, if $0 \leq H < 0.5$ the process is anti-correlated (for more information see e.g., Koutsoyiannis, 2010), for $0.5 < H \leq 1$ the process is positively correlated (most common case in geophysical processes) and for $H = 0.5$ the process is purely random (zero autocorrelation, thus white noise behaviour) at these large scales. Long-term persistence in natural processes was first discovered by H.E. Hurst (1951) while A. Kolmogorov (1941) mathematically described it, working on self-similar processes while studying turbulence. This behaviour is also known as the Hurst phenomenon or Hurst-Kolmogorov (HK) behavior (Koutsoyiannis, 2010). A stochastic process with HK behaviour with constant slope of climacogram ($-2 + 2H$) for all scales m (not only asymptotically), is known as a Hurst-Kolmogorov process or fractional Gaussian noise (see sect. 2 of the SM). In Table 1, we introduce the climacogram definition in case of a stochastic process in continuous time (eq. 2) and in discrete time (eq. 3), a widely used climacogram estimator (eq. 4) as well as climacogram estimation based on the latter estimator and expressed as a function of the true climacogram (eq. 5).

Table 1: Climacogram definition and expressions for a process in continuous and discrete time, along with the properties of its estimator.

Type	Climacogram	
continuous	$\gamma(m) := \dfrac{\mathrm{Var}\left[\int_t^{t+m}\underline{x}(\xi)\mathrm{d}\xi\right]}{m^2} = \mathrm{Var}\left[\int_0^m \underline{x}(\xi)\mathrm{d}\xi\right]/m^2$ where $m \in \mathbb{R}^+$ and $\gamma(0) \coloneqq \mathrm{Var}[\underline{x}(t)]$	(2)
discrete	$\gamma_\mathrm{d}^{(\Delta)}(k) := \frac{\mathrm{Var}\left[\sum_{l=k(i-1)+1}^{ki}\underline{x}_l^{(\Delta)}\right]}{k^2} = \frac{\mathrm{Var}\left[\sum_{l=1}^{k}\underline{x}_l^{(\Delta)}\right]}{k^2} = \gamma(k\Delta)$ where $k \in \mathbb{N}$ is the dimensionless scale for a discrete time process	(3)
classical estimator	$\hat{\underline{\gamma}}_\mathrm{d}^{(\Delta)}(k) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{k}\left(\sum_{l=k(i-1)+1}^{ki}\underline{x}_l^{(\Delta)}\right) - \frac{\sum_{l=1}^{n}\underline{x}_l^{(\Delta)}}{n}\right)^2$	(4)
expectation of classical estimator	$\mathrm{E}\left[\hat{\underline{\gamma}}_\mathrm{d}^{(\Delta)}(k)\right] = \frac{1-\gamma_d^{(\Delta)}(n)/\gamma_d^{(\Delta)}(k)}{1-k/n}\gamma_\mathrm{d}^{(\Delta)}(k)$	(5)

2.2 Autocovariance

The climacogram is fully determined if the autocovariance is known and vice versa. The specifics of the autocovariance, including its definition and estimator, are displayed in Table 2. Note that autocovariance is an even function.

Table 2: Autocovariance definition and expressions for a process in continuous and discrete time, along with the properties of its estimator.

Type	Autocovariance	
continuous*	$c(\tau) := \mathrm{Cov}[\underline{x}(t), \underline{x}(t+\tau)] = \dfrac{\mathrm{d}^2(\tau^2\gamma(\tau))}{2\mathrm{d}\tau^2}$ where $\tau \in \mathbb{R}$ is the lag for a continuous time process (in time units)	(6)
discrete	$c_\mathrm{d}^{(\Delta)}(j) := \mathrm{Cov}\left[\underline{x}_i^{(\Delta)}, \underline{x}_{i+j}^{(\Delta)}\right] = \dfrac{1}{2}\left((j+1)^2\gamma\left((j+1)\Delta\right) + (j-1)^2\gamma\left((j-1)\Delta\right) - 2j^2\gamma(j\Delta)\right)$ where $j \in \mathbb{Z}$ is the lag for the process at discrete time (dimensionless) and the right-hand side of the equation corresponds to the 2nd central finite derivative $j^2\,\gamma(j\Delta)$.	(7)
classical estimator	$\hat{\underline{c}}_\mathrm{d}^{(\Delta)}(j) = \frac{1}{\zeta(j)}\sum_{i=1}^{n-j}\left(\underline{x}_i^{(\Delta)} - \frac{1}{n}\left(\sum_{l=1}^{n}\underline{x}_l^{(\Delta)}\right)\right)\left(\underline{x}_{i+j}^{(\Delta)} - \frac{1}{n}\left(\sum_{l=1}^{n}\underline{x}_l^{(\Delta)}\right)\right)$ where $\zeta(j)$ is usually taken as: n or $n-1$ or $n-j$	(8)
expectation of classical estimator**	$\mathrm{E}\left[\hat{\underline{c}}_\mathrm{d}^{(\Delta)}(j)\right] = \dfrac{1}{\zeta(j)}\left((n-j)c_\mathrm{d}^{(\Delta)}(j) + \dfrac{j^2}{n}\gamma(j\Delta) - j\gamma(n\Delta) - \dfrac{(n-j)^2}{n}\gamma((n-j)\Delta)\right)$	(9)

*Eq. 6 can also be solved in terms of γ to yield (Koutsoyiannis 2013a): $\gamma(m) = 2\int_0^1(1-x)c(xm)\mathrm{d}x$.

**For proof see Appendix.

It is easy to see that for $\Delta > 0$:

$$c_{\mathrm{d}}^{(\Delta)}(0) := \gamma_{\mathrm{d}}^{(\Delta)}(1) = \gamma(\Delta) < \gamma(0) = c(0) \quad (10)$$

2.3 Power spectrum

Historically the power spectrum was defined in terms of the Fourier transform of the process $\underline{x}(t)$ by taking the expected value of the squared norm of the transform for time tending to infinity, which for a stationary process converges to the Fourier transform of its autocovariance (this is known as the Wiener- Khintchine theorem after Wiener, 1930, and Khintchine, 1934). Both definitions can be used for the power spectrum; however the latter is simpler and more operational and has been preferred in modern texts (e.g. Papoulis, 1991, ch. 12.4). In Table 3, we summarize the basic equations for the power spectrum definition and estimation.

Table 3: Power spectrum definition and expressions for a process in continuous and discrete time, along with the properties of its estimator.

Type	Power spectrum	
continuous*	$s(w) := 4 \int_0^\infty c(\tau) \cos(2\pi w \tau)\, \mathrm{d}\tau$ where $w \in \mathbb{R}$ is the frequency for a continuous time process (in inverse time units)	(11)
discrete**	$s_{\mathrm{d}}^{(\Delta)}(\omega) := 2\Delta\gamma(\Delta) + 4\Delta \sum_{j=1}^{\infty} c_{\mathrm{d}}^{(\Delta)}(j) \cos(2\pi\omega j)$ where $\omega \in \mathbb{R}$ is the frequency for a discrete time process (dimensionless; $\omega = w\Delta$)	(12)
classical estimator***	$\hat{\underline{s}}_{\mathrm{d}}^{(\Delta)}(\omega) = 2\Delta \hat{\underline{c}}_{\mathrm{d}}^{(\Delta)}(0) + 4\Delta \sum_{j=1}^{n} \hat{\underline{c}}_{\mathrm{d}}^{(\Delta)}(j) \cos(2\pi\omega j)$	(13)
expectation of classical estimator***	$\mathrm{E}\left[\hat{\underline{s}}_{\mathrm{d}}^{(\Delta)}(\omega)\right] = 2n\Delta\big(\gamma(\Delta) - \gamma(n\Delta)\big)/\zeta(0) + 4\Delta \sum_{j=1}^{n} \frac{\cos(2\pi\omega j)}{\zeta(j)} \left((n-j) c_{\mathrm{d}}^{(\Delta)}(j) + \frac{j^2}{n}\gamma(j\Delta) - j\gamma(n\Delta) - \frac{(n-j)^2}{n}\gamma((n-j)\Delta) \right)$	(14)

*Eq. 11 can be solved in terms of c to yield: $c(\tau) = \int_0^\infty s(w) \cos(2\pi w\tau)\, \mathrm{d}w$.

**Eq. 12 can be solved in terms of $c_{\mathrm{d}}^{(\Delta)}$ to yield: $c_{\mathrm{d}}^{(\Delta)}(j) = 1/\Delta \int_0^{1/2} s_{\mathrm{d}}^{(\Delta)}(\omega) \cos(2\pi\omega j)\, \mathrm{d}\omega$.

***Eq. 13 and 14 are more easily calculated with fast Fourier transform (fft) algorithms.

Note that power spectrum is an even function. As easily verified from eq. 12, in discrete time the power spectrum is periodic with period 1. Continuous and discrete time power spectra can be linked to each other by the simple equation (Koutsoyiannis, 2013a):

$$s_{\mathrm{d}}^{(\Delta)}(\omega) = \sum_{j=-\infty}^{\infty} s\left(\frac{\omega + j}{\Delta}\right) \sin^2\big(\pi(\omega + j)\big) / \{\pi(\omega + j)\}^2 \quad (15)$$

3. Statistical behaviour of the estimation of climacogram, autocovariance and power spectrum

Various physical interpretations of geophysical processes are based on the power spectrum and/or autocovariance behaviour (e.g. spectral density function of free isotropic turbulence, see in Pope, 2010, p. 610). However, the estimation of these tools from data may distort the true behaviour of the process and thus, may lead to wrong or unnecessarily complicated interpretation. To study the possible distortion we use the simplest processes often met in geophysics, which could also be used in synthesizing more complicated ones. Specifically, we investigate and compare the climacogram, autocovariance and power spectrum of various simple stochastic processes (whose expressions are presented in sect. 3.1) in terms of their behaviour and of their estimator performance (sect. 3.2 and 3.3) for different values of their parameters.

3.1 Testing stochastic models

To investigate the statistical behaviour of the estimators of the three tools, climacogram, autocovariance and power spectrum, we use two simple models. The first is the well-known Markovian model, else known as Ornstein-Uhlenbeck model, which has an exponentially decaying autocovariance. The second is a generalization of the HK process (abbreviation gHK), whose autocovariance decays as a power function of lag for large time lags while it is virtually an exponential function of lag, for small lags. Note that in sect. 2 of the SM, we also test the HK model.

In Table 4 and 5, we provide the mathematical expressions of the climacogram, autocovariance and power spectrum of a Markovian and gHK stochastic processes, respectively, in continuous and discrete time. Their estimates can be found from eq. 5, 9 and 14 and their model parameters, λ and q have dimensions $[x^2]$ and T, respectively, while b is dimensionless.

Table 4: Climacogram, autocovariance and power spectrum expressions of a Markovian process, in continuous and discrete time.

Type	Markovian process	
Autocovariance (continuous)	$c(\tau) = \lambda e^{-\lvert\tau\rvert/q}$	(16)
Autocovariance (discrete)	$c_d^{(\Delta)}(j) = \dfrac{\lambda\left(1 - e^{-\Delta/q}\right)^2}{(\Delta/q)^2}\, \mathrm{e}^{-(\lvert j\rvert-1)\Delta/q}$ for $\lvert j\rvert \geq 1$ and $c_d^{(\Delta)}(0) = \gamma(\Delta)$	(17)
Climacogram (for continuous and discrete)	$\gamma(m) = \dfrac{2\lambda}{(m/q)^2}\left(m/q + e^{-m/q} - 1\right)$ with $\gamma(0) = \lambda$	(18)
Power spectrum (continuous)	$s(w) = \dfrac{4\lambda q}{1 + 4\pi q^2 w^2}$	(19)
Power spectrum (discrete)	$s_d^{(\Delta)}(\omega) = 4\lambda q\left(1 - \dfrac{1}{\Delta/q}\,\dfrac{(1-\cos(2\pi\Delta\omega))\sinh(\Delta/q)}{\cosh(\Delta/q) - \cos(2\pi\Delta\omega)}\right)$	(20)

Table 5: Climacogram, autocovariance and power spectrum expressions of a positively correlated gHK process, with $0 < b < 1$, in continuous and discrete time.

Type	gHK process	
Autocovariance (continuous)	$c(\tau) = \lambda(\lvert\tau\rvert/q + 1)^{-b}$ with $b = 2 - 2H$	(21)
Autocovariance (discrete)	$c_d^{(\Delta)}(j) = \lambda \dfrac{\lvert j\Delta/q - \Delta/q + 1\rvert^{2-b} + \lvert j\Delta/q + \Delta/q + 1\rvert^{2-b} - 2\lvert j\Delta/q + 1\rvert^{2-b}}{(\Delta/q)^2(1-b)(2-b)}$ for $j \geq 1$, with $c_d^{(\Delta)}(0) = \gamma(\Delta)$	(22)
Climacogram (for continuous and discrete)	$\gamma(m) = \dfrac{2\lambda((m/q+1)^{2-b} - (2-b)m/q - 1)}{(1-b)(2-b)(m/q)^2}$ with $\gamma(0) = \lambda$	(23)
Power spectrum (continuous)	$s(w) \approx \dfrac{4\lambda q^b\,\Gamma(1-b)\mathrm{Sin}\left(\frac{\pi b}{2} + 2q\pi\lvert w\rvert\right)}{(2\pi\lvert w\rvert)^{1-b}} - \dfrac{4\lambda q\ {}_1\mathrm{F}_2\left[1; 1-\frac{b}{2}, \frac{3}{2}-\frac{b}{2}; -\pi^2q^2w^2\right]}{1-b}$ (where ${}_1\mathrm{F}_2$ is the hyper-geometric function)	(24)
Power spectrum (discrete) for q>0	not a closed expression*	-

* eq. 12 couldn't be further analysed

It should be noted that the gHK process can be considered as an HK process that gives a finite autocovariance value at zero lag, which is the common case in geophysical processes (an HK process with autocovariance $\lvert\tau\rvert^{-b}$ gives infinity at zero lag). Thus, a parameter q is added to the HK process indicating the limit between HK processes ($q \ll \lvert\tau\rvert$) and those affected by the minimum scale limit of the process ($q \gg \lvert\tau\rvert$). To switch to an HK process from the gHK one in the equations of Table 5, we can replace λ with λq^{-b} and then estimate the limit $q \to 0$ (see sect. 2.1 of the SM).

The expressions in Tables 4-5 are derived starting from the true autocovariance in continuous time (since most studies have preferred autocovariance-based computations; however the easiest way would be to start from the climacogram, to avoid the more complicated integral derived from eq. 6). Then, we can estimate its true value in discrete time and its expected value expressions (from eq. 7 and 9). Further, we can estimate the true values in continuous time as well as the expected values of the climacogram (from eq. 2 and 5) and finally, the true values in continuous and discrete time as well as the expected values of the power spectrum (from eq. 11, 12 and 14). From now on and for simplicity, only positive lags and frequencies will be considered as both the autocovariance and power spectrum are even functions.

3.2 Graphical investigation on the climacogram, autocovariance and power spectrum

We start our comparison with graphical investigations, which are actually very common in model identification. In Fig. 2-3, we have built the climacograms, autocovariances and power spectra for Markovian processes with q = 1, 10 and 100, and λ = 1 (Fig. 2) and gHK processes with q = 1, 10 and 100, b = 0.2 and $\lambda = q^{-b}$ (Fig. 3), all with $D = \Delta = 1$. In particular, in Fig. 2-3 we compare the true, continuous-time stochastic tools, along with their discrete-time versions as well as their expectation of classical estimators, as given in the equations of Tables 4-5. For the estimator, a medium sample size n = 10^3 was used (apparently, as n increases the bias will decrease). The graphs also contain plots of the negative logarithmic derivative (abbreviated as NLD) of all three functions. It is noted that the NLD is an important concept in identifying possible scaling behaviour (i.e. asymptotic power-laws like in the Hurst phenomenon) in geophysical processes and a useful metric for quantifying this behaviour (e.g., see Tyralis and Koutsoyiannis (2011) for the estimation of the Hurst coefficient). The NLD of any function $f(x)$ is defined as:

$$f^{\#}(x) ≔ -\frac{\mathrm{d}\ln(f(x))}{\mathrm{d}\ln x} = -\frac{x}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x} \tag{25}$$

and for the finite logarithmic derivative of $f(x)$, e.g. in case of discrete time process, we choose the forward logarithmic derivative, i.e.:

$$f^{\#}(x_{i+1}) ≔ -\frac{\ln(f(x_{i+1})/f(x_i))}{\ln(x_{i+1}/x_i)} \tag{26}$$

Figures 2-3 (including the analysis of the HK process in sect. 2.1 of the SM), allow us to make the following observations:

(a) As shown in eq. 3, the climacogram continuous-time values are equal to the discrete-time ones (for Δ = D > 0), while in case of the autocovariance and power spectrum they are different. More specifically, the discrete-time autocovariance ($c_{\mathrm{d}}^{(\Delta)}$) is practically indistinguishable from the continuous-time one (c), but only after the first lags, while the power spectrum continuous and discrete time values vary in both small and large frequencies (where this variation is larger in the latter).

(b) The expectation of autocovariance, $\mathrm{E}\left[\underline{\hat{c}}_{\mathrm{d}}^{(\Delta)}(j)\right]$, departs from both the true one (c) and the discrete-time one ($c_{\mathrm{d}}^{(\Delta)}$), for all the examined processes and its bias is always larger than that of the climacogram and the power spectrum (e.g., see also Lombardo et al., 2013). The climacogram has larger bias, in comparison with the power spectrum, in case of a gHK process (Fig. 3) and smaller bias for the Markovian one (Fig. 2).

(c) While in theory the NLD of the climacogram, autocovariance and power spectrum should correspond to each other, at least asymptotically (e.g., see Koutsoyiannis, 2013a), in practice, as observed in Fig. 2-3, this correspondence may be lost. In particular, on one hand, the NLDs of the discrete-time autocovariance (${c_d^{(\Delta)}}^{\#}$) and expectation value, $\mathrm{E}\left[\underline{\hat{c}}_d^{(\Delta)}(j)\right]^{\#}$, always tend to infinity in the high lag tail (due to the negative values produced). On the other hand, the NLD of the climacogram expectation value, $\mathrm{E}\left[\underline{\hat{\gamma}}^{(\Delta)}\right]^{\#}$, is close to the true one ($\gamma^{\#}$) for a Markovian process and increases with scale, in case of a gHK process. On the contrary, while for a Markovian process, the difference between the NLDs of the discrete-time power spectra (${s_d^{(\Delta)}}^{\#}$) and expectation value, $\mathrm{E}\left[\underline{\hat{s}}_d^{(\Delta)}(j)\right]^{\#}$, is small, in case of a gHK one, it is non-monotonic, as it varies in both low and high frequencies. Also, there is always a drop in the NLD of the power spectrum in the high frequency tail at ω = 0.5, which is attributed to the symmetry of the discrete-time and expectation of the power spectrum around ω = 0.5, leading to ${s_d^{(\Delta)}}^{\#}(0.5) = \mathrm{E}\left[\underline{\hat{s}}_d^{(\Delta)}(j)\right]^{\#}(0.5) = 0$.

(d) The expected value of the power can be estimated theoretically (through eq. 14) only up to frequency ω = 0.5 (which is the Nyquist frequency), due to the cosine periodicity. On the contrary, autocovariance and climacogram expected values can be estimated theoretically for scales and lags, respectively, up to n - 1.

(e) Finally, there is a high computational cost involved in the calculation of values and expectations of the power (taken from eq. 13 and 14, respectively) as compared to the simple expressions for the climacogram (eq. 5) and autocovariance (eq. 7 and 9), which is often dealt with fft algorithms. These large sums, along with the large number of trigonometric products, can often also cause numerical instabilities (e.g. in the gHK case, with q = 100, in Fig. 3e-f).

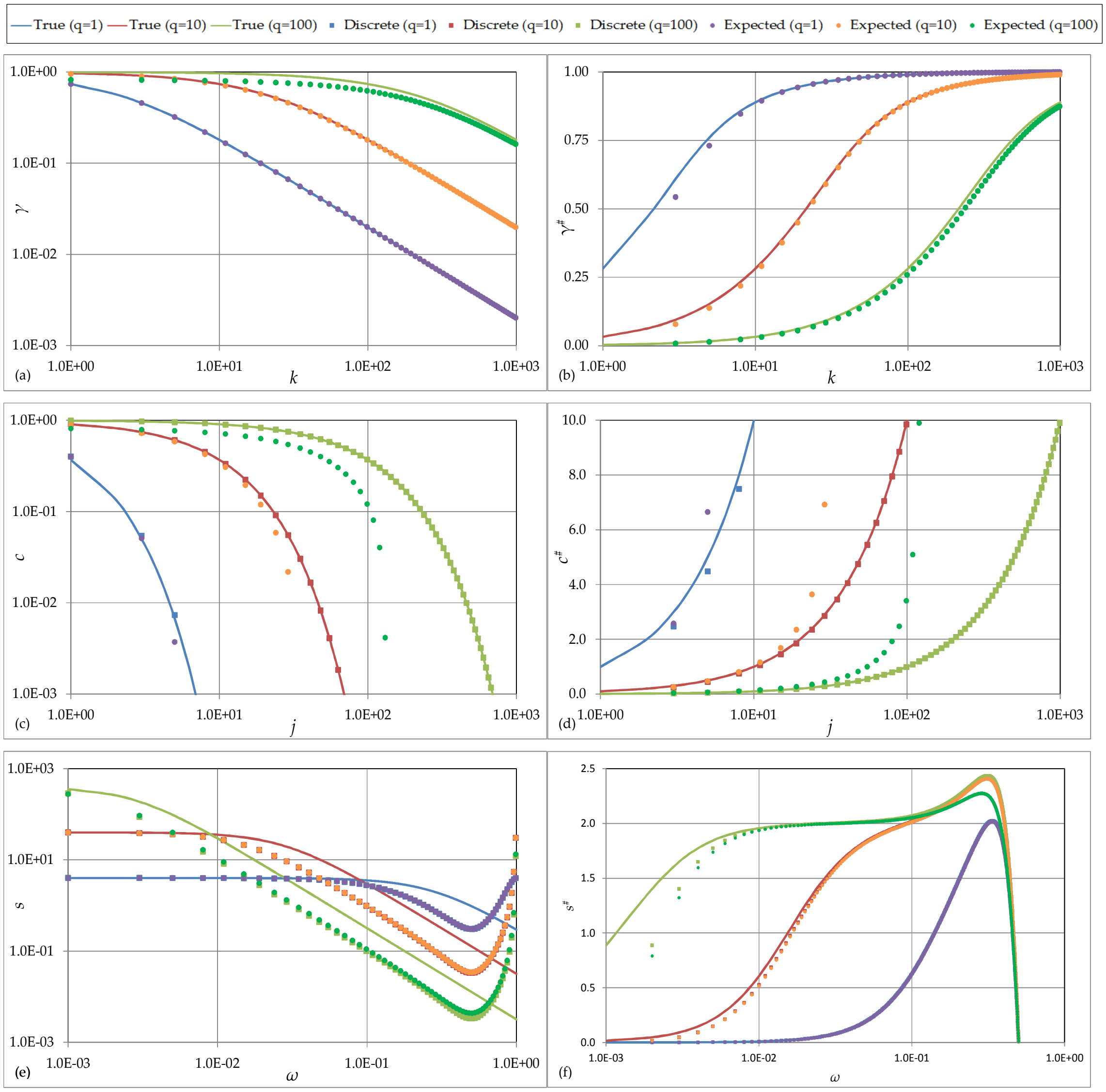

Figure 2: True values in continuous and discrete time and expected values of the climacograms (a), autocovariances (c) and power spectra (e) as well as their corresponding NLDs (b, d and f, respectively) of Markovian processes with q = 1, 10 and 100, λ = 1 and n = 10^3. Note that the continuous and discrete values of the climacogram are identical for $\Delta = D > 0$.

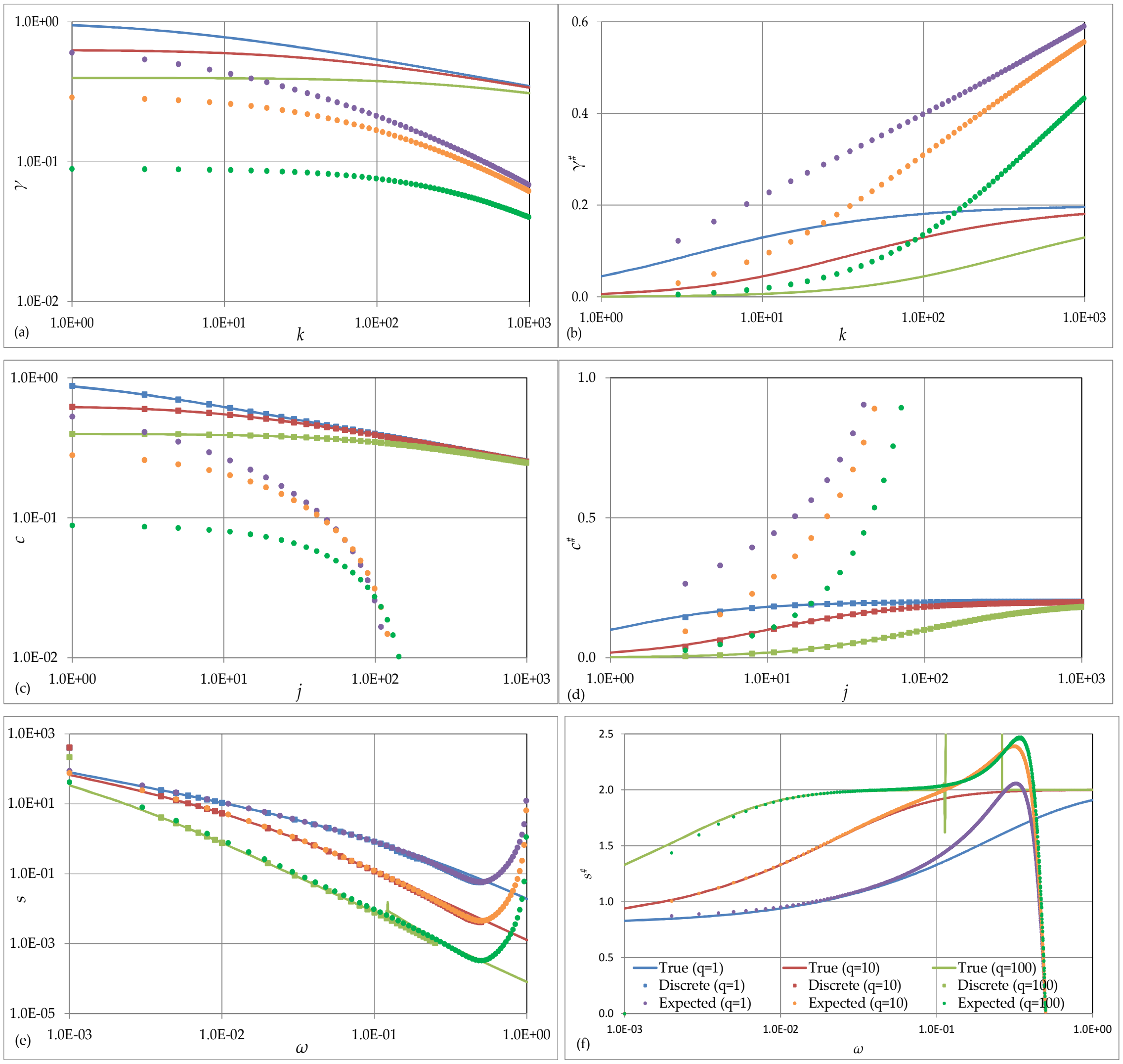

Figure 3: True values in continuous and discrete time and expected values of the climacograms (a), autocovariances (c) and power spectra (e) as well as their corresponding NLDs (b, d and f, respectively) of gHK processes with $b = 0.2$ and $q = 1$, 10 and 100, $\lambda = q^{-b}$ (not $\lambda = 1$, for demonstration purposes) and n=10^3. Note that the continuous and discrete values of the climacogram are identical for $\Delta = D > 0$.

Certainly, all the above are just indications arising from this graphical investigation of simple cases. For more complicated processes one should investigate further.

Some of the observations concerning the estimated power spectrum can be explained by considering the way the power spectrum is calculated from the autocovariance: when a sample value is above (below) the sample mean, the residual is positively (negatively) signed; thus, a high autocovariance value means that, in that lag, most of the residuals of the same sign are multiplied together (++ or --). In other words, the same signs are repeated (regardless of their difference in magnitude). The same 'battle of signs' process, is followed in the case of the power spectrum, but this time, the sign is given by the cosine function. A large value of the power spectrum indicates that, in that frequency, the autocovariance values multiplied by a positive sign (through the cosine function) are more than those multiplied by a negative one. So, the power spectrum can often misinterpret an intermediate change in the true autocovariance or climacogram. A way to track it down will be through the autocovariance itself, i.e. not using the power spectrum at all, but this is also prone to high bias (especially in its high lag tail) which always results in at least one negative value (for proof see Hassani, 2010 and analysis in Hassani, 2012). These can be avoided with an approach based on the climacogram, i.e. the variance of the time averaged process over averaging time scale, as the calculated variance is always positive. Also, the structure of the power spectrum is not only complicated to visualize and to calculate but also lacks direct physical meaning (opposite to autocovariance and climacogram), as it actually describes the Fourier transform of the autocovariance.

Furthermore, the power spectrum can often lead to process misinterpretations as the one shown in Fig. 2 (Markovian process), where almost in the whole frequency domain $\mathrm{E}\left[\underline{\hat{s}}_{\mathrm{d}}^{(\Delta)}\right] > s_{\mathrm{d}}^{(\Delta)}$ and $\left(s_{\mathrm{d}}^{(\Delta)}\right)^{\#} > \mathrm{E}\left[\underline{\hat{s}}_{\mathrm{d}}^{(\Delta)}(j)\right]^{\#}$. This can lead to the wrong conclusion that the area underneath $c_{\mathrm{d}}^{(\Delta)}$ is smaller than $\mathrm{E}\left[\underline{\hat{c}}_{\mathrm{d}}^{(\Delta)}\right]$ and that $c_{\mathrm{d}}^{(\Delta)}$ tends to zero more quickly than $\mathrm{E}\left[\underline{\hat{c}}_{\mathrm{d}}^{(\Delta)}\right]$. This can be easily derived from Fig. 4, if one replaces the cosine function with a simplified one (with only +1 and -1, where cosine is negative and positive, respectively). Then, the negative part of the simplified function lies with the negative part of the biased autocovariance, resulting in a positively signed value when multiplied with each other. However, this is not the case for the discrete autocovariance resulting in $\mathrm{E}\left[\underline{\hat{s}}_{\mathrm{d}}^{(\Delta)}\right] > s_{\mathrm{d}}^{(\Delta)}$.

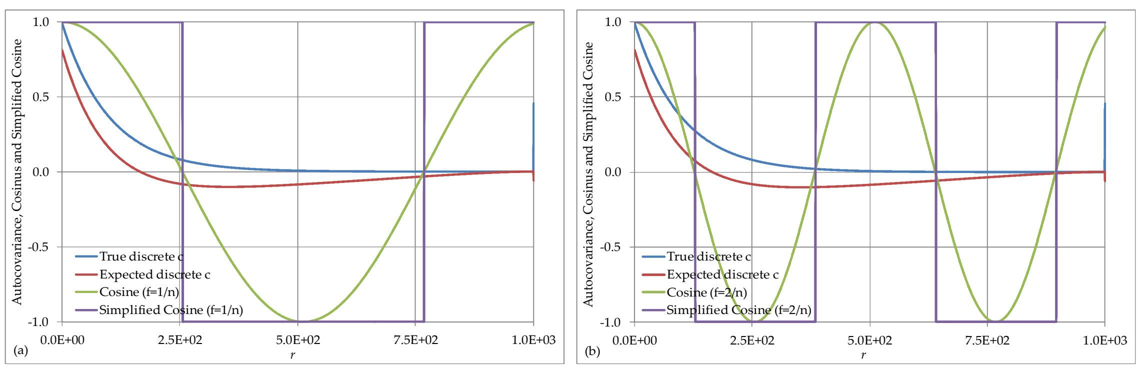

Figure 4: True autocovariance in discrete time for a Markov process (with q = 100) and its expected value for $n = 10^3$, along with a cosine function $\cos(2\pi f r)$, where f is the frequency and r the lag and its sign $\mathrm{sign}(\cos(2\pi f r))$, for (a) $f = 1/n$ and (b) $f = 2/n$.

3.3 Investigation of the estimators of climacogram, autocovariance and power spectrum

In this section, we will investigate the performance of the estimators of climacogram, autocovariance and power spectrum. For their evaluation we use mean square error expressions as shown in the equations below. Assuming that θ is the true value of a statistical characteristic (i.e. climacogram, autocovariance, power spectral density and NLDs thereof) of the process, a dimensionless mean square error (MSE), similar to the one used for the probability density function in Papalexiou et al. (2013), is:

$$\varepsilon = \frac{\mathrm{E}\left[\left(\underline{\hat{\theta}} - \theta\right)^2\right]}{\theta^2} = \varepsilon_v + \varepsilon_b \tag{27}$$

where we have decomposed the dimensionless MSE into a variance and a bias term, i.e.

$$\varepsilon_v = \mathrm{Var}\left[\underline{\hat{\theta}}\right]/\theta^2 \tag{28}$$

$$\varepsilon_b = \left(\theta - \mathrm{E}\left[\underline{\hat{\theta}}\right]\right)^2/\theta^2 \tag{29}$$

Note that θ is given by eq. 2 (for the true climacogram), eq. 7 (for the true autocovariance in discrete-time) and eq. 12 (for the true power spectrum in discrete-time). ε_b can be found analytically through $\mathrm{E}\left[\underline{\hat{\theta}}\right]$, from eq. 5, 9 and 14, respectively, but ε_v cannot (because of lack of analytical solutions for $\mathrm{E}\left[\underline{\hat{\theta}}^2\right]$ and hence, $\mathrm{Var}\left[\underline{\hat{\theta}}\right]$, for the classical estimators of climacogram, autocovariance and power spectrum). A way of tackling this would be by a Monte Carlo method, and specifically by producing many independent Gaussian synthetic time series with a known climacogram (and thus, autocovariance and power spectrum) and estimating the variance for each scale/lag/frequency, respectively. The methodology we used to produce synthetic time series, for any stochastic process based on a combination of Markovian processes (e.g., Mandelbrot, 1977), is given in sect. 3 of the SM. For a typical finite size n, the sum of a finite, usually small, number of Markovian processes is capable of adequate representing most processes; for example, Koutsoyiannis (2010) showed that the sum of 3 AR(1) models is adequate for representing an HK process for $n < 10^4$. Certainly, as accuracy requirements and n increase, a larger number of Markovian processes is required. Note that here, we do not use the AR(1) model to represent a process that is Markovian in continuous time (as shown in sect. 4 of the SM, the AR(1) model cannot represent a discretized continuous-time Markovian process for $\Delta/q > 0$ as well as $\Delta \neq D$). Instead, we use the ARMA(1,1) model which (as mentioned in Koutsoyiannis, 2002, 2013a) successfully represents any Markovian process and in sect. 4 of the SM we derive its parameters.

Thus, we produce synthetic time series for Markovian processes with q = 1, 10 and 100 (Fig. 5) and gHK ones with q = 1, 10 and 100 and b = 0.2 (Fig. 6), all with $D = \Delta = 1$. Then, for each scale, lag and frequency, we calculate for all processes the means, variances, means of the NLD, and variances of the NLD, for the climacogram, autocovariance and power spectrum, and their corresponding errors through eq. 27 to 29, for $n = 10^3$ (Fig. 5-6) and for $n = 10^2$ and 10^4 (sect. 2.2 of the SM). Note that, on one hand, as n decreases, both bias and variance increase and thus, for the point estimate and variance to be closer to the expected ones, we need more time series. On the other hand, as n increases, more Markovian processes have to be added and with a larger bias and variance (due to larger q). So, for the examined processes, we conclude that in order to achieve a maximum error of about 1‰ between scales 1 and $n/2$, we have to produce approximate 10^4 time series for $n = 10^2$, 10^3 and 10^4. The error is meant here as the absolute difference, between the estimated and expected value, divided by the expected value. Furthermore, the 1‰ error refers to the climacogram and corresponds to a gHK process with b = 0.2 and q = 100, which is considered the more adverse of the examined processes. Note that in Fig. 5-6, we try to show all estimates within a single plot for comparison with each other. The inverse frequency in the horizontal axis is set to $1/(2\omega)$, so as to vary between 1 and $n/2$ and the lag to j+1, so as the estimation of variance at $j = 0$ is also shown in a log-log plot.

Moreover, we investigate the shape of the probability density function (pdf) for each stochastic tool, which, in many cases, differs from a Gaussian one, resulting in deviations between the mean (expected) and mode. To measure this difference, we use the sample skewness (denoted g), where for $g \approx 0$, the difference is small and for any other case, larger. In Fig. 7, we show for each stochastic tool and for a gHK process with b = 0.2 and q/Δ = 10, an example of their 95% upper and lower confidence intervals (corresponding to exceedence probabilities of 2.5% and 97.5%), as well as their pdf for a specific scale, lag and frequency.

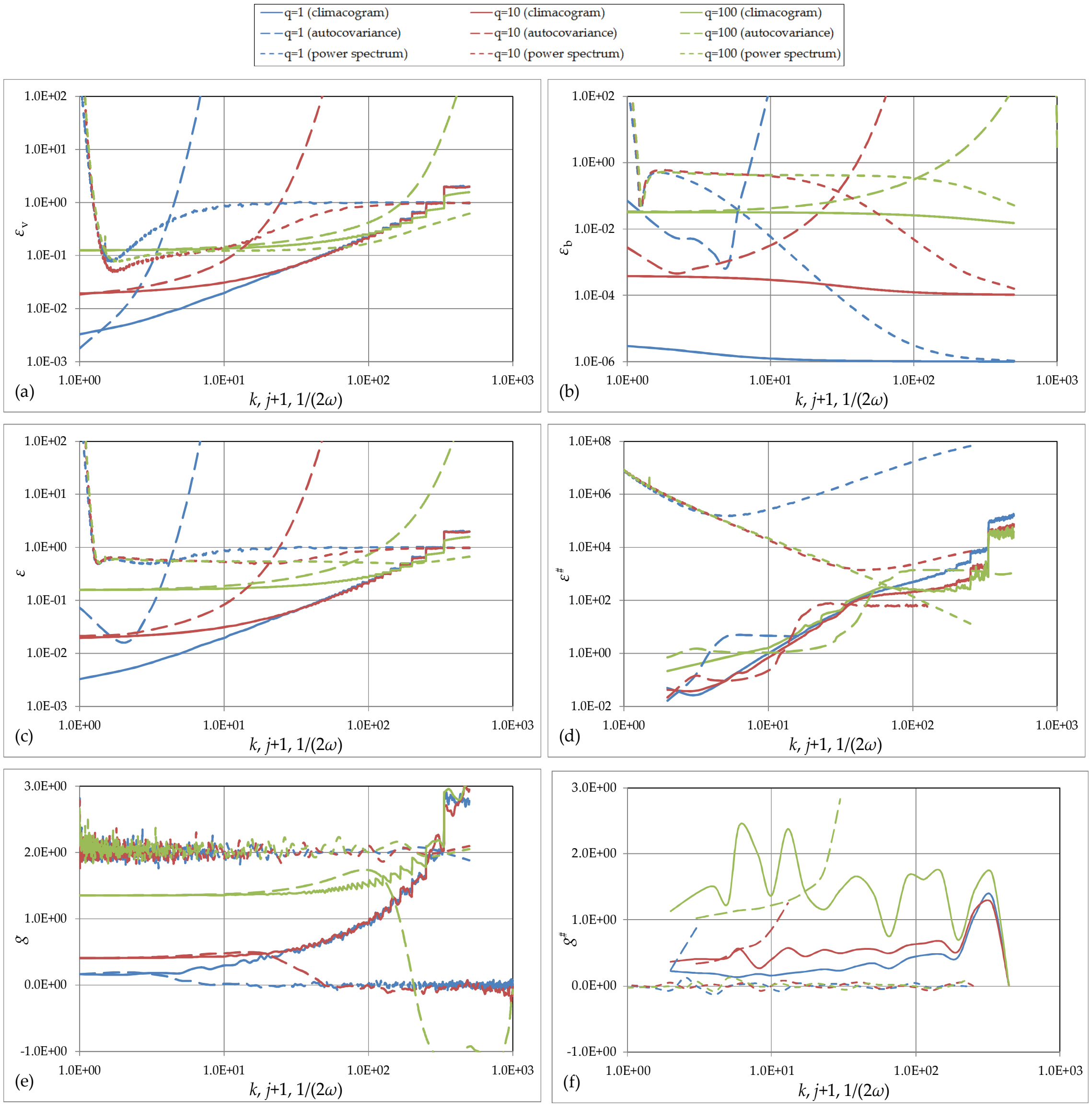

Figure 5: Dimensionless errors of the climacogram estimator (continuous line), autocovariance (dashed line) and power spectrum (dotted line), calculated from 10^4 Markovian synthetic series with $n = 10^3$ (for $b = 0.2$, $q = 1$, 10 and 100 and $\lambda = q^{-b}$): (a) ε_{v} (dimensionless MSE of variance); (b) ε_{b} (dimensionless MSE of bias); (c) ε (total dimensionless MSE); and (d) $\varepsilon^{\#}$ (total dimensionless MSE of NLD); as well as the sample skewness of each of the stochastic tools and their NLDs are also shown (e) and (f).

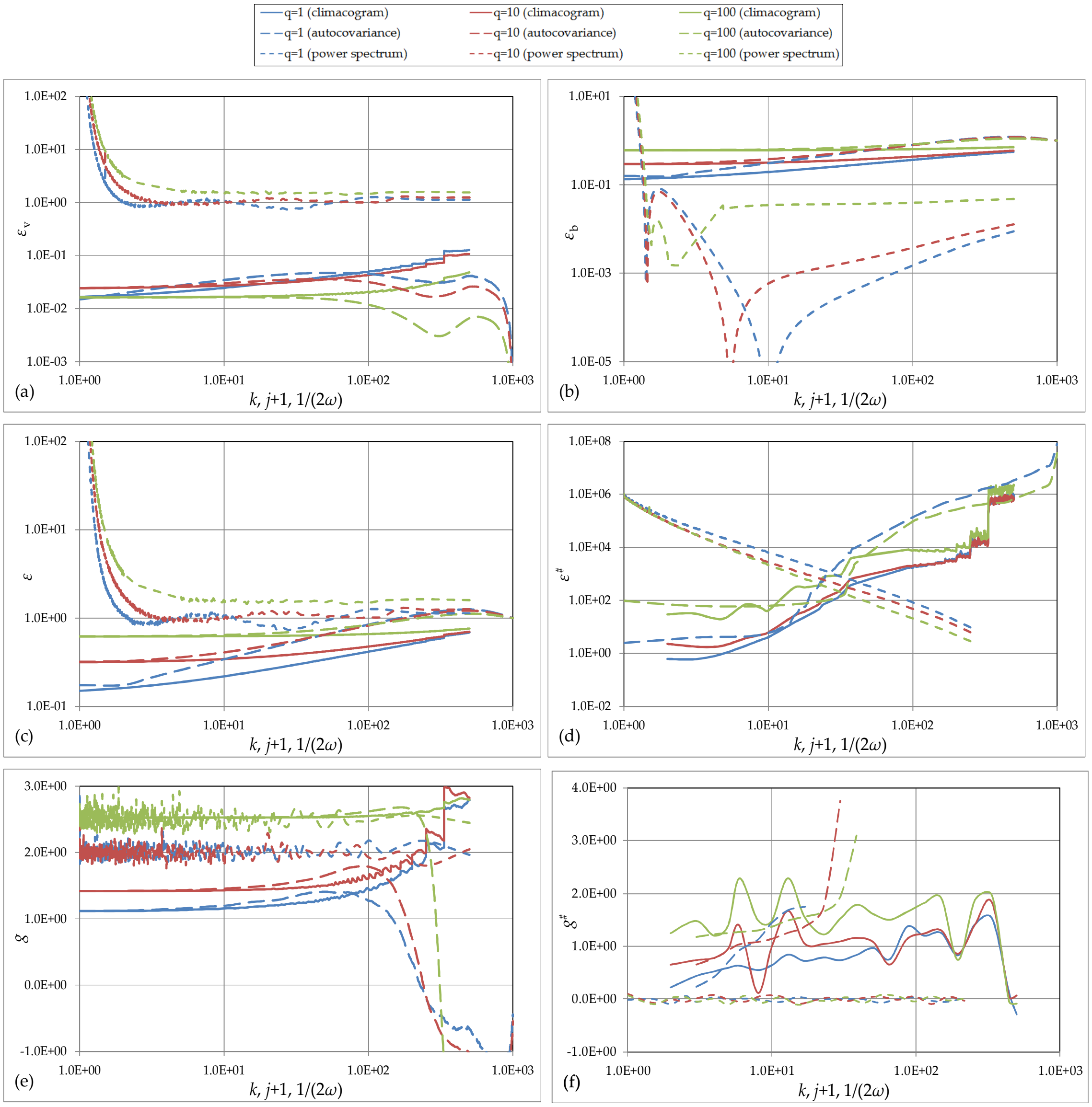

Figure 6: Dimensionless errors of the climacogram estimator (continuous line), autocovariance (dashed line) and power spectrum (dotted line), calculated from 10^4 gHK synthetic series with n = 10^3 (for b = 0.2, q = 1, 10 and 100 and $\lambda = q^{-b}$): (a) ε_{v} (dimensionless MSE of variance); (b) ε_{b} (dimensionless MSE of bias); (c) ε (total dimensionless MSE); and (d) $\varepsilon^{\#}$ (total dimensionless MSE of NLD); as well as the sample skewness of each of the stochastic tools and their NLDs are also shown in (e) and (f).

Figures 5-6 (including the analysis in sect. 2.2 of the SM), allow us to make some observations related to stochastic model building:

(1) In general, the climacogram has lower variance than that of the autocovariance, which in turn is lower than that of the power spectrum (e.g. Markovian and HK processes as well as gHK for most scales). Also, it has a smaller bias than that of the autocovariance but larger than the one of the power spectrum (for all examined processes). Since, for the Markovian and HK processes, the error component related to the variance, ε_{v}, is usually larger than the one from the bias, ε_{b}, or conversely for the gHK ones, the climacogram has a smaller total error ε, in most cases. Thus, we can state that (for all the examined cases) the expression below holds:

$$\mathrm{E}\left[\left(\underline{\hat{\gamma}}-\gamma\right)^2\right]/\gamma^2 \leq \mathrm{E}\left[\left(\underline{\hat{c}}_d^{(\Delta)}-c_d^{(\Delta)}\right)^2\right]/{c_d^{(\Delta)}}^2 \leq \mathrm{E}\left[\left(\underline{\hat{s}}_d^{(\Delta)}-s_d^{(\Delta)}\right)^2\right]/{s_d^{(\Delta)}}^2 \tag{30}$$

(2) We see that as n and b (for the HK process) or q (for the Markovian and gHK processes) increase, the climacogram estimator entails much smaller error than that of the autocovariance and power spectrum for the whole domain of scales, lags and frequencies.

(3) The total error for the NLD, $\varepsilon^\#$, increases with scale in the climacogram and with lag in the autocovariance for all examined processes. In case of an exponentially decaying autocovariance (e.g. in a Markovian process), the power spectrum slope $\varepsilon^\#$ first decreases and then increases in large inverse-frequency values, while the autocovariance and climacogram $\varepsilon^\#$ always increase. In this type of process, climacogram and autocovariance $\varepsilon^\#$ are close to each other and in most cases smaller than the power spectrum $\varepsilon^\#$. For HK and gHK processes, where large scales/lags/inverse-frequencies exhibit an HK behaviour, the power spectrum always decreases with inverse frequency under a power-law decay, in contrast to the autocovariance and climacogram $\varepsilon^\#$ which they always increase. Thus, in this type of processes, there exists a cross point between power spectrum $\varepsilon^\#$ and the other two, where behind this point, the power spectrum has a larger $\varepsilon^\#$ and beyond a smaller one.

(4) The pdf of the climacogram and autocovariance have small skewness magnitude and can approximate a Gaussian pdf for most of scales and lags, while the power spectrum pdf has a larger skewness for its regular values (besides its theoretical smaller bias), which results in non-symmetric confidence intervals (very important when it comes to uncertainty in stochastic modeling, e.g., see Lombardo et al., 2014). However, the NLD of the power spectrum has a negligible skewness in comparison with those of the autocovariance and climacogram, which means that the expected NLD should be very close to the NLD mode.

(5) The climacogram skewness is increasing with scale up to 3, while the autocovariance one is larger at first and then it drops to -1 (the point where it starts to drop is when the expected autocovariance reach a negative value for the first time). It is interesting that the power spectrum skewness has a value around 2 for regular values and 0 for NLDs, for all the examined processes (with the exception of the extreme gHK process with $q/\Delta = 100$, where it is around 2.5).

(6) The power spectrum has a large ε in high frequencies and then it stabilizes around 1 for all the examined processes and n. This observation is also mathematically verified by Papoulis (1991, p. 449, eq. 13-59). Also, we observe that the autocovariance and climacogram ε always increases with scale and lag, respectively.

(7) The autocovariance ε is decreasing with q, for the examined Markovian and gHK processes, and increasing with b for the examined HK ones. In contrast, the climacogram ε is increasing with q (for the examined Markovian and gHK processes) and decreasing with b (for the HK ones).

(8) The autocovariance and power spectrum $\varepsilon^\#$ are decreasing with q, for the examined Markovian and gHK processes, and increasing with b for the examined HK ones. The climacogram $\varepsilon^\#$ is decreasing with both q and b.

(9) The climacogram exhibits sudden increases of ε and $\varepsilon^\#$ (like a stairway) beyond scales equal to the 10%-20% of $n/2$ (maximum possible scale for the climacogram). This is due to the small number of data from which the variance is calculated. This is also verified by Koutsoyiannis (2003, 2013a) leading to a rule of thumb of estimating the climacogram until the $n/10$ (20% of $n/2$) scale.

(10) $\mathrm{Var}\left[\underline{\hat{s}}^\#\right]$ has a power-type decay with inverse-frequency with an exponent around -2.0 to -2.5, for all the examined processes.

(11) We observe that the variance of the power spectrum, for all the examined processes and sample sizes, is approximately equal to the square of its expected value for frequencies $\omega \neq 0$, 0.5 and 1 and double the square of its expected value for $\omega = 0$, 0.5 and 1. This is also verified by Papoulis (1991, p. 447, eq. 13-50) and discussed in Press et al. (2007, p. 655).

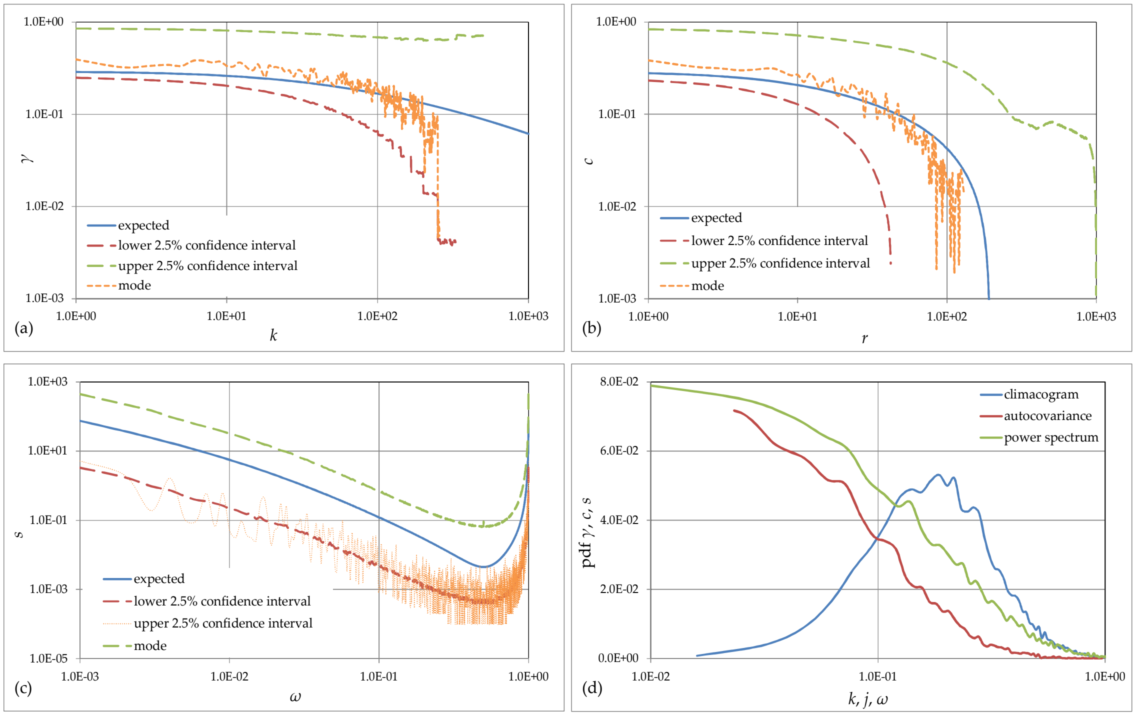

Figure 7: Expected value (continuous blue line), upper 95% confidence interval (dashed green line), lower 95% confidence interval (dashed red line) and mode for (a) climacogram, (b) autocovariance and (c) power spectrum and (d) climacogram empirical pdf (blue line), autocovariance (red line) and power spectrum (green line), at $k = j = 100$ and $\omega = 0.1$, respectively, calculated from 10^4 gHK synthetic series, with $b = 0.2$, q=10, $\lambda = q^{-b}$ and $n = 10^3$.

Apparently, these results are valid for the simple processes examined, and the typical estimator and sample sizes used, while to draw conclusions for more complex processes, the above analyses should be repeated. On the one hand, we can conclude that from observations 1 and 2, it is more likely for the sample climacogram to be closer to the theoretical one (considering also the bias) in comparison to the sample autocovariance or power spectrum to be closer to their theoretical values. Thus, it is proposed to use the climacogram when building a stochastic model and estimate the autocovariance and power spectrum from that model, rather than directly from the data (see application in sect. 4). On the other hand, it seems from observation 3, that in case of a power-law decay in large scales, lags and inverse-frequencies (e.g. in a HK or a gHK process) the NLD of that decay (i.e. b which is related to the HK coefficient) is better estimated from the power spectrum rather than the climacogram or autocovariance. However, this applies only for inverse-frequencies beyond the cross point (discussed in the 3rd observation). This can be tricky as we do not know where this point lies and also, this rule doesn't apply for exponential autocovariance decay (e.g. in a Markovian process) where the NLD is now very large and again, it can lead to wrong conclusions about the nature of the large scale decay (i.e. presence or not of the Hurst phenomenon). In conclusion, the observations 1-3 can be used to build a general frame of rules of thumb (described in the steps below) to build a stochastic model from a sample or to interpret its physical process, e.g. identify what type of process is (Markovian, HK, gHK etc.). This framework is based only on the three examined stochastic tools and it should be expanded in case more tools are to be used in the analysis. An application to a real-world example is presented in sect. 4 for illustration purposes.

(a) First, we have to decide upon the large scale type of decay from the climacogram. For example, if the large scale NLD is close to 1 then the process is more likely to exhibit either an exponential decay of autocovariance at large lags (scenario S1) or a white noise behaviour, i.e. $H = 0.5$ (scenario S2). In case where the large scale NLD deviates from 1 then the process is more likely to exhibit an HK behaviour (scenario S3). The autocovariance can help us choose between scenarios S1 and S2, as in S1 we expect an immediate, exponential-like, drop of the autocovariance (which often has the smaller difference between its expected and mode value) whereas in S2 it is unbiased and therefore, the NLD should be close to 1. In case of the scenario S1, we can estimate the scale parameter of the Markovian-type decay from the NLD of the climacogram while in case of S3, we should also look into the power spectrum decay behaviour in low frequencies. Thereafter, for the determination of the Hurst coefficient, we can use various algorithms, e.g., the one of Tyralis and Koutsoyiannis (2010), which is based on the climacogram (usually taken up to 10%-20% of its maximum scale $n/2$), or that of Chen et al. (2010), which is based on the power spectrum.

(b) For the estimation of the rest of the properties (e.g. for intermediate and smaller scales) we should use the climacogram.

(c) To build a model, we should first try to use a combination of the processes used in this paper, i.e. an combination of Markovian, HK and gHK processes, as they are the simplest ones (principle of parsimony), with an immediate physical interpretation and their combination should cover most of the cases. If they do not represent well the physical process, we can use more complicated mathematical processes but repeating for each one the graphical investigation and statistical analysis proposed in this paper (for example, as done in sections 3.2 and 3.3).

(d) After we built our model, we should make the statistical analysis proposed in section 3.3, to verify our initial assumptions (null hypothesis) on the smaller ε and $\varepsilon^{\#}$ of the process as well as their pdf skewness magnitude, concerning its climacogram, autocovariance and power spectrum.

4. Application

In this section, we will show a statistical analysis of a set of 40 time series derived from a large open access dataset (http://www.me.jhu.edu/meneveau/datasets/datamap.html), provided by the Johns Hopkins University, which consists of turbulent wind velocity data, measured by X-wire probes downstream of an active grid at the direction of the flow (Kang et al., 2003). The first 16 time series correspond to velocities measured at transverse points abstaining $r = 20M$ from the source, where M = 0.152 m is the size of the grid placed at the source. The next 4 time series correspond to a distance r = $30M$, the next 4 to $40M$ and the last 16 to $48M$ (for more details concerning the experimental setup and data, see Kang et al., 2003). We have chosen this type of dataset for our application because of the controlled environment of the experiment, as well as for its broad importance as turbulence drives almost any geophysical process. Additionally, all time series have a nearly-Gaussian probability density function (see Fig. 8b) and are nearly isotropic (isotropy ratio 1.5, see in Kang et al., 2003). Also, their sample sizes are very large, $n = 10^6$ data for each time series (the original data set consisted of 36×10^6 data values but, following Koutsoyiannis (2012) approach, we averaged every 36 observations, resulting in 10^6 observations, for the sake of simplicity). Yet D remains small (0.9 ms for the averaged time series) and thus, the equality $D \approx \Delta$ can still be assumed valid. Finally, the data set gives the opportunity of cross checking the methodology proposed in section 3.3, by applying it firstly for the averaged process (Fig. 9a-d) derived from all 40 time series and then for a single one (Fig. 9d) with statistical characteristics close to the averaged one. In all cases stationarity is assumed, given that the macroscopic flow characteristics are steady. The modelling of higher moments and derivatives of the process, which are important for phenomena such as intermittency and bottleneck effects, as well as interpretation of model parameters, is not within the scope of this paper. We only focus on the preservation of the 2nd order statistics related to the three examined stochastic tools.

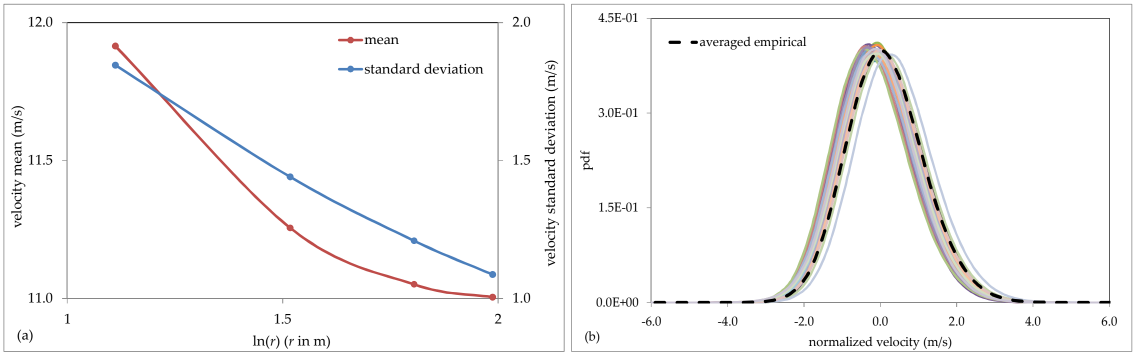

Figure 8: Data preliminary analysis: a) averaged velocity mean (red line) and averaged standard deviation (blue line) along the wind tunnel axis and (b) empirical pdfs of the normalized time series (by subtracting the mean and dividing with the standard deviation, for each time series) and their averaged empirical pdf (black thick line).

In Fig. 9, we show the climacograms, autocovariances and power spectra of all the 40 normalized time series, their averaged values and the corresponding values for the 38th time series whose stochastic properties are closest to the averaged one. We choose to analyze this single time series to show a comparison with the averaged one. Notice here, that we do not apply the windowing technique to eliminate some of the power spectrum variance as it is causing loss of information for small frequencies (see Fig. 10d). Also, windowing should be used with caution when choosing small segment lengths and should be avoided in strongly correlated processes (e.g. the ones that exhibit Hurst behaviour) as the time series of the divided segments are not independent from each other.

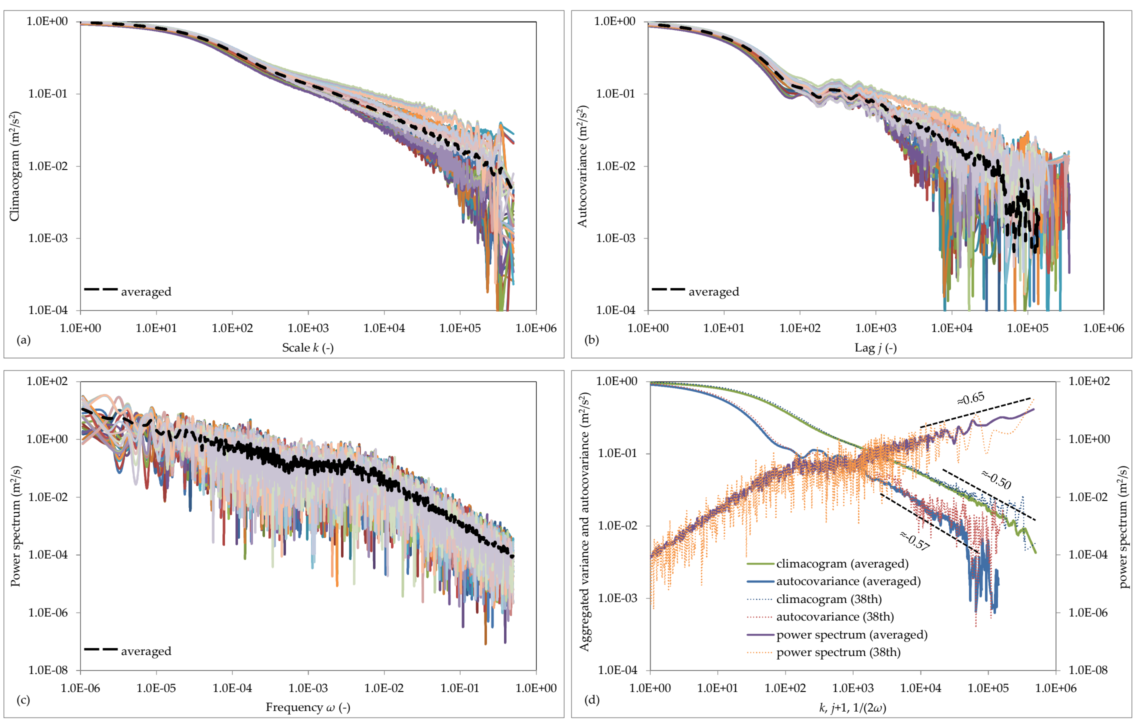

Figure 9: Data stochastic analysis: (a) climacograms, (b) autocovariances and (c) power spectra of all the 40 time series (multi-coloured lines) as well as their averaged values (dashed thick black line), (d) all three in one plot focusing on the comparison of the averaged values with those of the 38th time series; NLDs at large scales, lags and inverse frequencies are also shown.

The velocity field is not homogeneous in the direction of the flow, e.g. the velocity mean and standard deviation in every position is decreasing with the distance r from the source as shown in Fig. 8a. To homogenize all time series, we normalize each one by subtracting the mean (red line) and dividing with the standard deviation (blue line).

Assuming that the averaged values, shown in Fig. 9d, are close to the expected values of the process, we can fit a model following the proposed methodology in section 3.3. The large scale NLD is far from 1, hence, it is most likely that the process exhibits a Hurst behaviour, i.e. a power law decay of the autocovariance (scenario S3). For the identification of the process' behaviour at intermediate and small scales, we use the climacogram as it is more likely to have the least standardized variance (as shown in sect. 3.3). We finally observe that the NLD at small scales can be very well represented by a Markovian process. Thus, we fit a stochastic model consistent with the observed behaviour (as seen on the climacogram) combining Markovian and gHK processes. Namely, we fit a model (Table 6) consisting of one Markovian process (controlling small scale behaviour) and a gHK process (controlling large scale behaviour).

Table 6: Autocovariance, climacogram and power spectrum mathematical expressions, in continuous and discrete time, of a composite model consisted of a Markovian and a gHK process.

Type	Stochastic model	
Autocovariance (continuous)	$c(\tau) = \lambda_1 e^{-\lvert\tau\rvert/q_1} + \lambda_2(\lvert\tau\rvert/q_2 + 1)^{-b}$	(31)
Autocovariance (discrete)	$c_d^{(\Delta)}(j) = \dfrac{\lambda_1\left(1 - e^{-\Delta/q_1}\right)^2}{(\Delta/q_1)^2}\, \mathrm{e}^{-(\lvert j\rvert-1)\Delta/q_1} + \lambda_2 \dfrac{\lvert j\Delta/q_2 - \Delta/q_2 + 1\rvert^{2-b} + \lvert j\Delta/q_2 + \Delta/q_2 + 1\rvert^{2-b} - 2\lvert j\Delta/q_2 + 1\rvert^{2-b}}{(\Delta/q_2)^2(1-b)(2-b)}$ with $c_d^{(\Delta)}(0) = \gamma(\Delta)$	(32)
Climacogram (continuous and discrete)	$\gamma(m) = \dfrac{2\lambda_1}{(m/q_1)^2}\left(m/q_1 + e^{-m/q_1} - 1\right) + \dfrac{2\lambda_2\{(m/q_2+1)^{2-b} - (2-b)m/q_2 - 1\}}{(1-b)(2-b)(m/q_2)^2}$ with $\gamma(0) = \lambda_1 + \lambda_2$	(33)
Power spectrum (continuous)	$s(w) \approx \dfrac{4\lambda_1 q_1}{1 + 4\pi q_1{}^2 w^2} + \dfrac{4\lambda_2 q_2{}^b\,\Gamma(1-b)\mathrm{Sin}\left(\frac{\pi b}{2} + 2q_2\pi\lvert w\rvert\right)}{(2\pi\lvert w\rvert)^{1-b}} - \dfrac{4\lambda_2 q_2\ {}_1\mathrm{F}_2\left[1; 1-\frac{b}{2}, \frac{3}{2}-\frac{b}{2}; -\pi^2 q^2 w^2\right]}{1-b}$	(34)
Power spectrum (discrete)	not a closed expression (see in Table 5)	-

As a first priority, we try to best fit the climacogram of the time series and on a secondary basis, the autocovariance and power spectrum (see Fig. 10). To estimate the parameters of the model two alternative fitting errors were considered:

$$\varepsilon_{s\gamma} = \sum_{k=1}^{n/2}\left\{\frac{\mathrm{E}\left[\underline{\hat{\gamma}}(\Delta k)\right] - \underline{\hat{\gamma}}_d^{(\Delta)}(k)}{\mathrm{E}\left[\underline{\hat{\gamma}}(\Delta k)\right]}\right\}^2 \tag{35}$$

$$\varepsilon_{\mathrm{m}\gamma} = \max_{k=1,\dots,n/2}\left|\frac{\mathrm{E}\left[\underline{\hat{\gamma}}(\Delta k)\right]-\underline{\hat{\gamma}}_d^{(\Delta)}(k)}{\mathrm{E}\left[\underline{\hat{\gamma}}(\Delta k)\right]}\right| \tag{36}$$

where $\underline{\hat{\gamma}}_d^{(\Delta)}(k)$ is the empirical climacogram (estimated from data) and $\mathrm{E}\left[\underline{\hat{\gamma}}(\Delta k)\right]$ the expected one (estimated from the model). Firstly, we use the $\varepsilon_{\mathrm{s}\gamma}$ error to locate initial values and then the $\varepsilon_{\mathrm{m}\gamma}$ for fine tuning and distributing the error equally to all scales. The optimization analysis results in: λ_1=0.81 and λ_2 = 0.19 m²/s², q_1 = 0.504 ms and q_2 = 5.04 ms and b = 0.45 (H=0.775), with $\varepsilon_{\mathrm{m}\gamma}$ = 41% and the R^2 equal to ~100% for the climacogram, 99.9% for the autocovariance and 99.0% for the power spectrum.

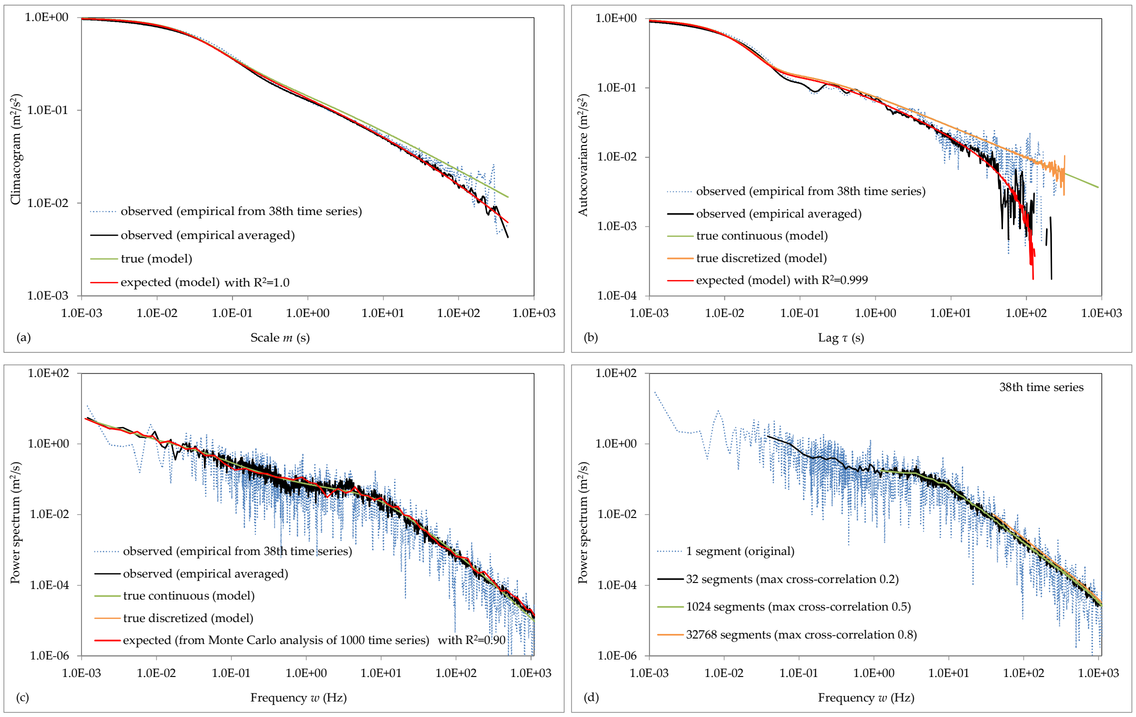

Figure 10: (a) Climacogram, (b) autocovariance and (c) power spectrum for the model of Table 6 fitted to turbulence data: true values in continuous time (estimated from the model – shown with a green line), true values in discrete time (estimated from the model – shown with an orange line), expected values (estimated from the model – shown with a red line), empirical averaged (estimated from all 40 time series – shown with a purple line) and sample values (estimated from the 38[th] time series – shown with a dashed blue line). Note that, to avoid large computational burden, the expected values of the power spectrum are not calculated from eq. 14, but from a Monte Carlo analysis of 10^4 synthetic time series. In (d) Bartlett's method is applied for the 38[th] time series for various numbers of segments and the cross-correlation between segments is shown.

Note that in Fig. 10d, Bartlett's method (Welch method for non-overlapping segments and with the use of a uniform window) is applied for the 38[th] time series. The increase of the cross-correlation with the increase of the number of segments the original time series is divided into, causes an increase to the dependence between segments, and thus, highlights the inappropriateness of this method in estimating the expected power spectrum. Finally, to test the validity of our assumption that for the specific model in Table 6, the estimator based on the climacogram has the smallest error ε compared to those based on the autocovariance and power spectrum, we use the same analysis proposed in step (d) in section 3.3. We produce 10^4 time series with n = 10^6 and we compare the errors ε for each estimator for 81 points logarithmically distributed from 1 to n (Fig. 11). Following the methodology of sect. 3 and 4 of the SM, we fit the gHK process in Table 6 with 7 Markovian models, with: p_1 = 26.622, p_2 = 6.

377 and $\varepsilon_{rm} \approx 0.2\%$. As can be observed from Fig. 11, the initial choice of the climacogram based estimators to identify the true process from the sample (null hypothesis), is proven valid for the current model and for all examined scales (in comparison with the other two estimators). Specifically, for all time scales the climacogram is more skillful for the estimation of both regular and NLD values of the process. The only clear exceptions are the smallest magnitude of the sample skewness of the autocovariance in the last lags and those of the NLD of the power spectrum (which means that their pdfs are closer to Gaussian and thus, their mode value is closer to their mean). However, these advantages are diminished by their larger variance and/or bias related errors. Here, it is also observed that the power spectrum errors seem to be quite constant not only for ε (as expected from the analysis in sect. 3.3) but for $\varepsilon^{\#}$ as well. This is due to the mixing of increasing Markovian process $\varepsilon^{\#}$ (see Fig. 4) and to the decreasing power-type ones (see Fig. 6 for the gHK process). The larger fluctuations of the power spectrum, in contrast to the climacogram and autocovariance ones, in Fig. 11, are indicative of its larger statistical variance and thus, of the smaller likelihood that the empirical power spectrum is closer to the expected one from the model.

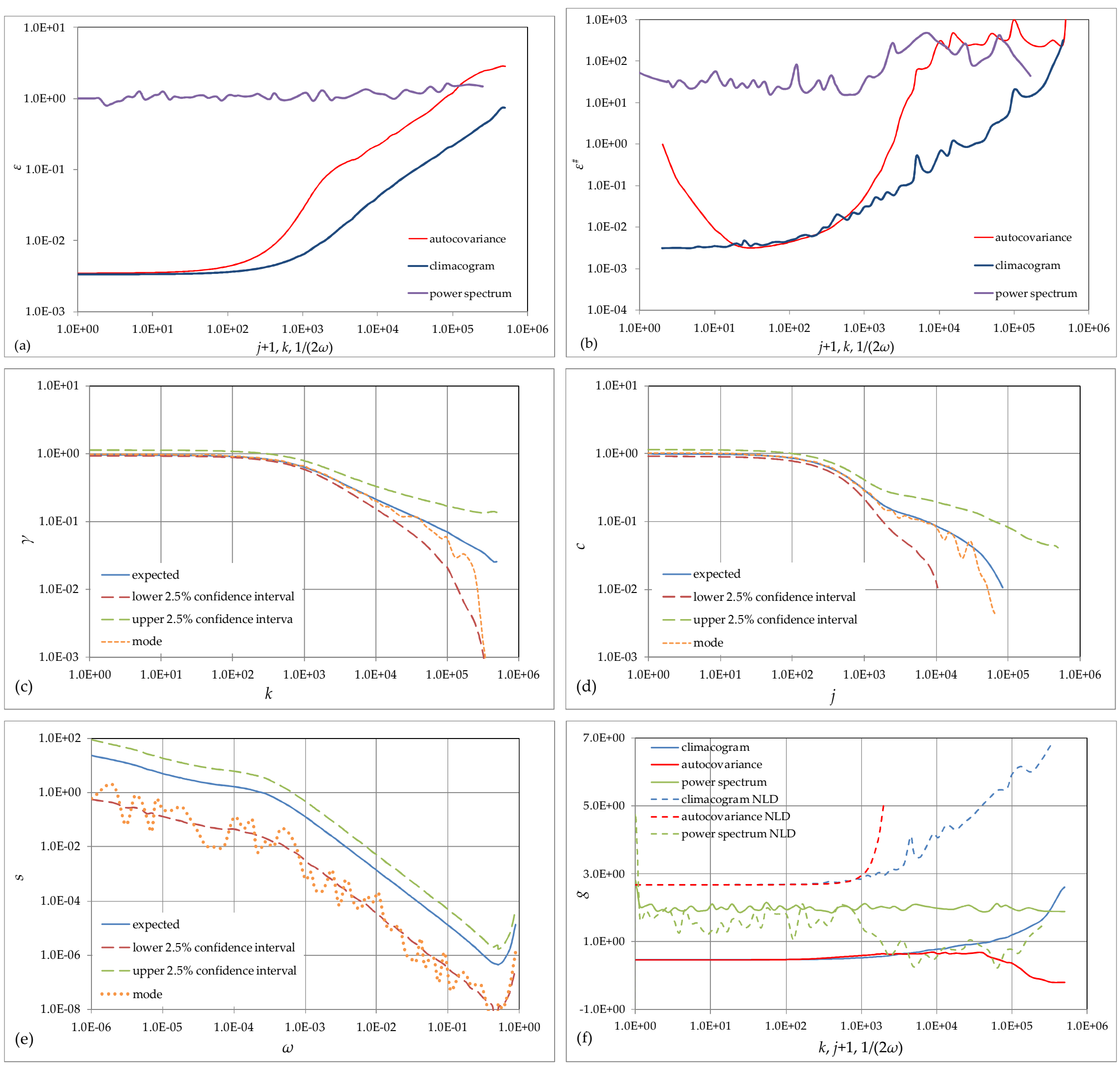

Figure 11: Dimensionless errors (a) ε and (b) $\varepsilon^{\#}$ of the climacogram and autocovariance compared with the power spectrum, as well as their expected values, along with upper and lower 95% confidence intervals and mode (c, d and e), as well as (f) skewness, calculated from 10^4 synthetic series with $n = 10^6$ based on the process in Table 6.

5. Summary and conclusions

The applications of the autocovariance and power spectrum, in order to identify the stochastic structure of natural processes and construct models thereof, abound in the literature. Less frequent is the use of the climacogram, which is a simpler tool and is related by one-to-one transformation to both the autocovariance and power spectrum. However, in very few cases the estimation uncertainty and bias are included in the calculations, causing possible inconsistencies and misspecifications of the model sought. Here we provide a theoretical framework to calculate the uncertainty and bias for those three stochastic tools, which also enables inter-comparison of the three tools and identification of their advantages and disadvantages.

For the climacogram and the autocovariance, analytical formulae for the calculation of the bias are possible and are presented here; in particular, the expected value of the classical estimator of the autocovariance in terms of the true climacogram and true autocovariance in discrete time is derived here (Eq. 9 and Appendix) and it is shown how it can be decomposed into only four parts, which are easy to evaluate. In contrast, the power spectrum, due to its more complicated definition (based on the Fourier transform of the autocovariance), does not enable a generic, analytically derived, formula for the estimation bias.

The study shows some of the advantages and difficulties presented in stochastic model building when starting from the climacogram, autocovariance or power spectrum. Specifically:

- The climacogram has the smallest estimation error in estimating the true values (in all examined cases as described in eq. 30) as well as the true logarithmic derivatives, i.e. slopes in log-log plots (with few exceptions). Also, its bias can be estimated through a simple and analytical expression (Eq. 5). Moreover, the climacogram is always positive (a property helpful in stochastic model building, e.g. the logarithmic derivative always exists), well-defined (with an intuitive definition through the variance of the time averaged process over averaging time scale) and typically monotonic (observed in all the examined processes and in the NLDs, in Fig. 2-3 and in sect. 2.1 of the SM). Finally, it has (for all the examined processes) values of sample skewness close to 0, for the small scale tail, while in the large scale tail, its skewness is increasing up to 3 (Fig. 5-6 and sect. 2.2 of the SM).
- The autocovariance has estimation errors larger than those of the climacogram. Besides its large bias, it is also prone to discretization errors as its value (eq. 7) can never be equal with the true value in continuous time (eq. 6), even for an infinite sample size. Moreover, it has negative values in the high lag tail (creating difficulties in stochastic model building, e.g. the logarithmic derivative does not exist). However, it is well-defined (with an intuitive definition), and with the help of Eq. 9, its bias can be estimated through a simple and analytical expression. Finally, it has (for all the examined processes) values of skewness close to 0, for the small lag tail, while in the large lag tail, its skewness is decreasing down to -1 (Fig. 5-6 and sect. 2.2 of the SM).
- The power spectrum has the largest values of estimation error (in all examined cases it is mostly around 100% of the true value in discrete time). Besides its bias, it is also prone to discretization error as its value (eq. 12) can never be equal to the true value in continuous time (eq. 11) even for an infinite sample size. Moreover, while theoretically its values are positive, numerical calculations based on data can result in negative values. In addition, it has a complicated definition (based on the Fourier transform of the autocovariance), which also involves complicated and high computational cost calculations for the discrete time and expected values (eq. 12-14 and Fig. 6e-f), as well as a non-monotonic NLD (observed in all the examined processes; Fig. 2-3 and sect. 2.1 of the SM). Finally, it often has the highest value of skewness for its regular values (mostly constant around 2) and the smallest one (around 0) for its NLD values (Fig. 5-6 and sect. 2.2 of the SM). The latter advantage of the power spectrum means that its mode should be close to the expected one, which however, is difficult to estimate, due to the aforementioned reasons.

The above theoretical and experimental results allow us to draw a general conclusion that the climacogram could provide a more direct, easy and accurate means both to make diagnoses from data and build stochastic models in comparison to the power spectrum and autocovariance.

As incidental contributions of the paper, we mention in the SM (sect. 3) the proposed methodology to produce synthetic Gaussian distributed time series of a process by decomposing it in multiple Markovian processes. This methodology is based only on an equation providing the scale parameters of the Markovian processes. Furthermore, we developed an ARMA(1,1) model in the SM (sect. 4), appropriate for simulating discrete-time Markovian processes; the need to introduce this, is related to the fact that the errors produced by a discrete-time AR(1) model (whose equivalent continuous-time process exhibits Markovian properties only when Δ = 0) when Δ > 0, can be significant for large first-order autocorrelation coefficient (see Fig. 5 of the SM).

Acknowledgement

This paper was partly funded by the Greek General Secretariat for Research and Technology through the research project "Combined REnewable Systems for Sustainable ENergy DevelOpment" (CRESSENDO; programme ARISTEIA II; grant number 5145). We thank the anonymous Associate Editor and the three anonymous Reviewers for the constructive comments which helped us to improve the paper, as well as the Springer Correction Team for editing the manuscript.

References

Chen, Y., R. Sun and A. Zhou (2010), An improved Hurst parameter estimator based on fractional Fourier transform, *Telecommunication Systems*, 43(3/4), 197–206.

Dimitriadis P., D. Koutsoyiannis and Y. Markonis (2012), Spectrum vs Climacogram, *European Geosciences Union General Assembly 2012*, Geophysical Research Abstracts, Vienna, Session HS7.5/NP8.3: Hydroclimatic stochastics, EGU2012-993.

Fleming S.W. (2008), Approximate record length constraints for experimental identification of dynamical fractals, *Ann. Phys.* (Berlin) 17, No. 12, 955-969.

Fourier J. (1822), Théorie analytique de la chaleur, *Firmin Didot Père et Fils*, Paris.

Gilgen, H. J. (2006), Univariate time series in geosciences: Theory and examples, *Berlin: Springer*.

Hassani H. (2010), A note on the sum of the sample autocorrelation function, *Physica A* 389, 1601-1606.

Hassani H. (2012), The sample autocorrelation function and the detection of long-memory processes, *Physica A* 391, 6367-6379.

Hurst, H.E., 1951. Long term storage capacities of reservoirs, *Trans. Am. Soc. Civil Engrs*., 116, 776–808.

Kang H.S., S. Chester and C. Meneveau (2003), Decaying turbulence in an active-grid-generated flow and comparisons with large-eddy simulation, *J. Fluid Mech.* 480, p. 129-160.

Khintchine, A. (1934), Korrelationstheorie der stationären stochastischen Prozesse, *Mathematische Annalen*, 109(1): 604–615.

Kolmogorov, A.N., 1941. Dissipation energy in locally isotropic turbulence, *Dokl. Akad. Nauk. SSSR*, 32, 16-18.

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

Koutsoyiannis, D. (2003), Climate change, the Hurst phenomenon, and hydrological statistics, Hydrological Sciences Journal, 48 (1), 3–24.

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

Koutsoyiannis, D. (2012), Re-establishing the link of hydrology with engineering, *Invited lecture at the National Institute of Agronomy of Tunis (INAT)*, Tunis, Tunisia.

Koutsoyiannis, D. (2013a), Encolpion of stochastics: Fundamentals of stochastic processes, 12 pages, *Department of Water Resources and Environmental Engineering – National Technical University of Athens*, Athens.

Koutsoyiannis, D. (2013b), Climacogram-based pseudospectrum: a simple tool to assess scaling properties, *European Geosciences Union General Assembly 2013*, Geophysical Research Abstracts, Vol. 15, Vienna, EGU2013-4209, European Geosciences Union.

Lombardo, F., E. Volpi and D. Koutsoyiannis (2013), Effect of time discretization and finite record length on continuous-time stochastic properties, *IAHS - IAPSO - IASPEI Joint Assembly*, Gothenburg, Sweden, International Association of Hydrological Sciences, International Association for the Physical Sciences of the Oceans, International Association of Seismology and Physics of the Earth's Interior.

Lombardo, F., E. Volpi, S. Papalexiou and D. Koutsoyiannis (2014), Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology, *Hydrol. Earth Syst. Sci*.

Mandelbrot, B. B. (1977), The Fractal Geometry of Nature, *Freeman, New York, USA*.

Papalexiou, S.M., D. Koutsoyiannis and C. Makropoulos (2013), How extreme is extreme? An assessment of daily rainfall distribution tails, *Hydrology and Earth System Sciences*.

Papoulis A. (1991), Probability, Random Variables and Stochastic Processes, 3rd edition, *McGraw Hill*.

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

Pope, S.B., 2000. Turbulent Flows, *Cambridge University Press*.

Stoica P. and Moses R. (2004), Spectral Analysis of Samples, *Prentice Hall*.

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

Wiener N. (1930), Generalized Harmonic Analysis, *Acta Mathematica*, 55: 117–258.

Appendix

Here, we express the expected value of the discrete time autocovariance in terms only of its true continuous time value using the corresponding true climacogram. This is very useful in stochastic modelling as it saves computational time (compared to a direct calculation where a sum throughout all the discrete time autocovariances is needed) and also because it gives a physical interpretation of the expected discrete time autocovariance.

Eq. 2 can be expressed in terms of the true discrete autocovariance:

$$\gamma(k\Delta) = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} c_d{}^{(\Delta)}(i-j) = \frac{2}{k^2}\sum_{i=1}^{k-1}(k-i)c_d{}^{(\Delta)}(i) + \frac{\gamma(\Delta)}{k} \tag{37}$$

The estimation of autocovariance in eq. 9 can be analysed to:

$$\mathrm{E}\left[\underline{\hat{c}}_d^{(\Delta)}(j)\right] = \mathrm{E}\left[\frac{1}{\zeta(j)}\sum_{i=1}^{n-j}\left(\underline{\hat{x}}_i^{(\Delta)} - \frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right)\right)\left(\underline{\hat{x}}_{i+j}^{(\Delta)} - \frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right)\right)\right] = \frac{1}{\zeta(j)}\sum_{i=1}^{n-j}\mathrm{E}\left[\left(\left(\underline{\hat{x}}_i^{(\Delta)} - \mu\right) - \left(\frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right) - \mu\right)\right)\left(\left(\underline{\hat{x}}_{i+j}^{(\Delta)} - \mu\right) - \left(\frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right) - \mu\right)\right)\right] = \frac{1}{\zeta(j)}\sum_{i=1}^{n-j}\left\{\overbrace{\mathrm{E}\left[\left(\underline{\hat{x}}_i^{(\Delta)} - \mu\right)\left(\underline{\hat{x}}_{i+j}^{(\Delta)} - \mu\right)\right]}^{\mathrm{E1}} - \overbrace{\mathrm{E}\left[\left(\underline{\hat{x}}_i^{(\Delta)} - \mu\right)\left(\frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right) - \mu\right)\right]}^{\mathrm{E2}} - \overbrace{\mathrm{E}\left[\left(\underline{\hat{x}}_{i+j}^{(\Delta)} - \mu\right)\left(\frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right) - \mu\right)\right]}^{\mathrm{E3}} + \overbrace{\mathrm{E}\left[\left(\frac{1}{n}\left(\sum_{l=1}^{n}\underline{\hat{x}}_l^{(\Delta)}\right) - \mu\right)^2\right]}^{\mathrm{E4}}\right\} \tag{38}$$

where $\mu = \mathrm{E}\left[\underline{\hat{x}}_i^{(\Delta)}\right]$.

Below we will express the above sums of expressions E1, E2, E3 and E4 in terms of the true climacogram $\gamma(\Delta k)$ and true autocovariance in discrete time $c_d{}^{(\Delta)}(j)$ for $j \geq 1$. Firstly, the sum of E1 is:

$$\sum_{i=1}^{n-j}\mathrm{E1} = (n-j)c_d{}^{(\Delta)}(j) \tag{39}$$

We observe that $\sum_{i=1}^{n-j}\mathrm{E2} = \sum_{i=1}^{n-j}\mathrm{E3}$ and thus, we only calculate the sum of E3:

$$\sum_{i=1}^{n-j} \mathrm{E3} = \frac{1}{n}\sum_{i=1}^{n-j}\sum_{l=1}^{n} \mathrm{E}\left[\left(\underline{\hat{x}}_{i+j}^{(\Delta)} - \mu\right)\left(\underline{\hat{x}}_{l}^{(\Delta)} - \mu\right)\right] = \frac{1}{n}\sum_{i=1}^{n-j}\sum_{l=1}^{n} c_d{}^{(\Delta)}(l-i-j) = \frac{(n-j)^2\gamma(\Delta(n-j))}{n} + \frac{1}{n}\overbrace{\sum_{i=1}^{n-j}\sum_{l=1}^{j} c_d{}^{(\Delta)}(l-i-j)}^{\mathrm{E5}} \tag{40}$$

The sum of E4 can be expressed in terms of the true climacogram:

$$\sum_{i=1}^{n-j} \mathrm{E4} = (n-j)\,\mathrm{Var}\left[\frac{1}{n}\left(\sum_{l=1}^{n} \underline{\hat{x}}_{l}^{(\Delta)}\right)\right] = (n-j)\,\gamma(n\Delta) \tag{41}$$

For the estimation of E5, we distinguish two cases, $j \le n/2$ and $j > n/2$. For the first case, we have:

$$\mathrm{E5}\left(j \le \frac{n}{2}\right) = \frac{j}{n}\sum_{i=j}^{n-j} c_d{}^{(\Delta)}(i) + \frac{1}{n}\sum_{i=1}^{j-1} i c_d{}^{(\Delta)}(i) + \frac{1}{n}\sum_{i=n-j+1}^{n-1} c_d{}^{(\Delta)}(i)(n-i) =$$

$$= \frac{j\gamma(\Delta) - j^2\gamma(j\Delta)}{2n} + \frac{1}{n}\overbrace{\sum_{i=n-j+1}^{n-1} c_d{}^{(\Delta)}(i)(n-i) + j\sum_{i=1}^{n-j} c_d{}^{(\Delta)}(i)}^{\mathrm{E6}} \tag{42}$$

For the estimation of E6, we have:

$$\mathrm{E6} = n\gamma(n\Delta)/2 - \gamma(\Delta)/2 + \frac{1}{n}\overbrace{\sum_{i=1}^{n-j} c_d{}^{(\Delta)}(j-n+i)}^{\mathrm{E7}} \tag{43}$$

and E7 can be expressed as:

$$\mathrm{E7} = (n-j)\gamma(\Delta)/(2n) - (n-j)^2\gamma\big((n-j)\Delta\big)/(2\mathrm{n}) \tag{44}$$

For $j > n/2$, E5 is the same as for $j \le n/2$ but with replacing j with n-j and thus, in the general case of E5:

$$\mathrm{E5} = n\gamma(n\Delta)/2 - j^2\gamma(j\Delta)/(2n) - (n-j)^2\gamma\big((n-j)\Delta\big)/(2n) \tag{45}$$

Thus, eq. 38 results in:

$$\mathrm{E}\left[\underline{\hat{c}}_d^{(\Delta)}(j)\right] = \frac{1}{\zeta(j)}\left((n-j)c_d{}^{(\Delta)}(j) + \frac{j^2}{n}\gamma(j\Delta) - j\gamma(n\Delta) - \frac{(n-j)^2}{n}\gamma\big((n-j)\Delta\big)\right) \tag{46}$$

where $\zeta(j)$ is usually taken as: n or $n-1$ or $n-j$.

It is interesting to notice that using eq. 7 we can express the expected discrete time autocovariance of the above using only the true climacogram.