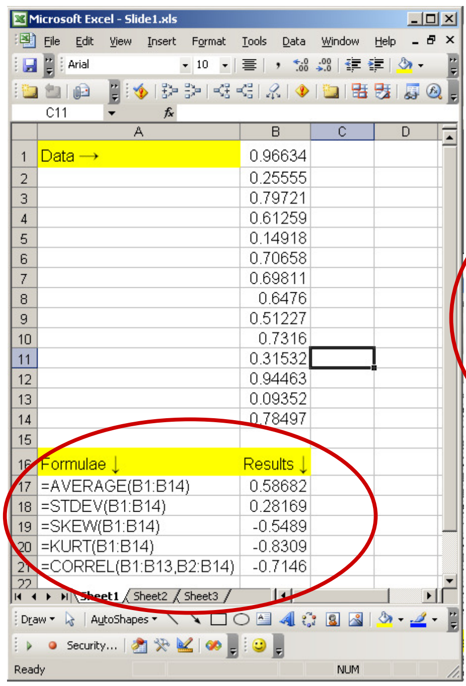
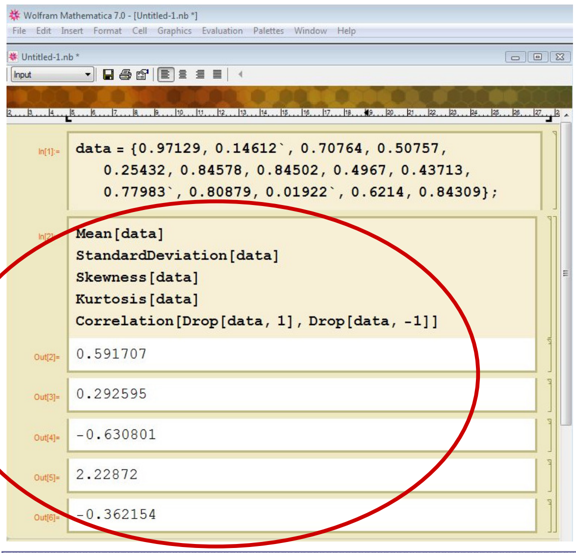
International Association of Hydrological Sciences
STAHY Official Workshop: Advances in statistical hydrology
Taormina, Italy, 23 - 25 May 2010

Mind the bias!

S.M. Papalexiou, D. Koutsoyiannis
Department of Water Resources and Environmental Engineering, National Technical University of Athens, Greece
A. Montanari
Faculty of Engineering, University of Bologna, Italy

Introductory note: Stochastics is more than calculations





Popular computer programs have made calculations easy and fast
But numerical results may mean **nothing**!
It is better not to use them if we are unaware of the stochastic properties of the objects

Misuse case 1: When bias is theoretically zero

- Experiment: A Google search with terms *multifractal rainfall moments* was performed (see also Koutsoyiannis, 2010)
- The first (highest PageRank) paper was chosen and its first figure is reproduced here

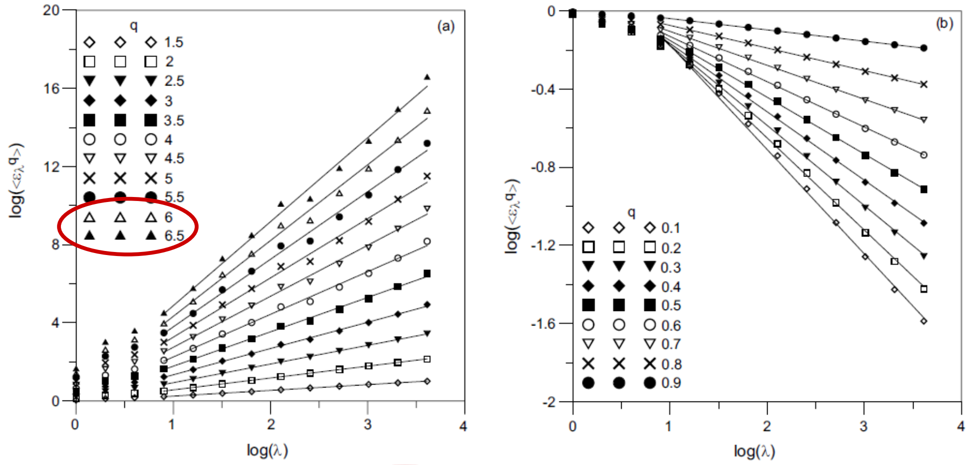


Fig. 1. Log-log plot of the qth moments of the rainfall intensity on the time scales from 1 hour to almost 6 months versus the scale ratio λ. (a) For moments larger than 1; (b) for moments smaller than 1.

Can we really calculate the high moments of rainfall depths?

- High moments, i.e. $m_q := \mathrm{E}[x^q]$ for q = 4, 5, 6, 7, ..., depend enormously and exclusively on the distribution tail
- Recent research results (e.g. Koutsoyiannis 2004, 2005; Papalexiou and Koutsoyiannis, 2010; and references therein) suggest power-type/Pareto tail with shape parameter κ = 0.13-0.15, almost constant worldwide
- This reflects the (imperfect) **scaling in state** of rainfall rate
- Beyond $q_{max} = 1/\kappa = 6.67$ (for $\kappa = 0.15$) the moments are infinite
- However, their numerical estimates from a time series are always finite: an **infinite negative bias**
- But below q_{max} it can be proved that the estimates are **unbiased**
- However, even below q_{max}, the estimation of moments is problematic; this can be demonstrated by Monte Carlo simulation

Setting up the Monte Carlo (MC) simulation

- Random variable $\underline{x}$ (representing rainfall distribution tail, i.e. rainfall excess above a certain threshold)
- Pareto distribution function with parameters κ (shape) and λ (scale)

  $P\{\underline{x} > x\} =: F^*(x) = (1 + \kappa\, x/\lambda)^{-1/\kappa}$
- Analytically calculated moments ($B(\ )$ denotes the beta function)

  $m_q = \mathrm{E}[\underline{x}^q] = q\,(\lambda/\kappa)^q\, B(1/\kappa - q,\, q)$ for $q < 1/\kappa$

  $m_q = \mathrm{E}[\underline{x}^q] = \infty$ for $q \geq 1/\kappa$
- Random sample $\underline{x}_1, \underline{x}_2, \ldots \underline{x}_n$, with size $n = 100$
- Moment estimator (a random variable)

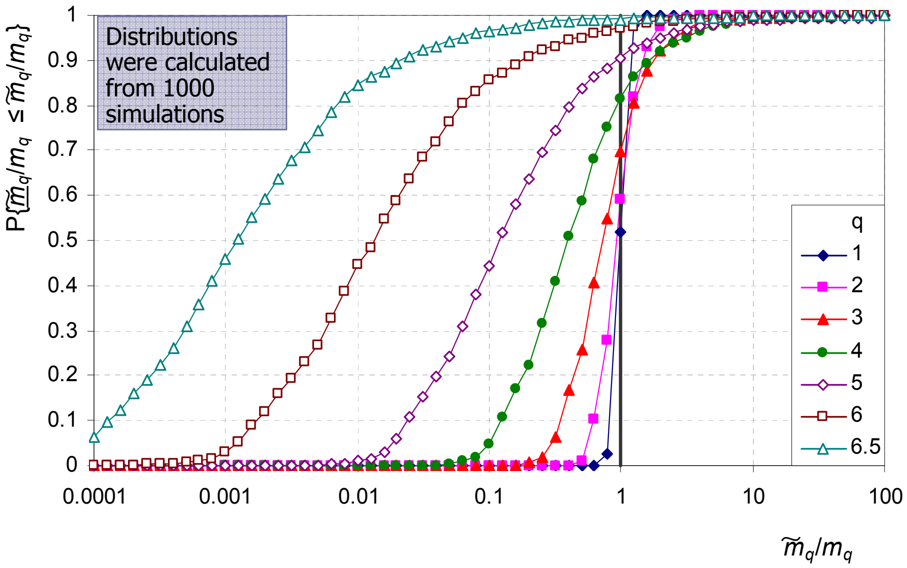
  $\underline{\tilde{m}}_q = (1/n) \sum_{i=1}^{n} \underline{x}_i^q$ Note: $\mathrm{E}[\underline{\tilde{m}}_q] = m_q \rightarrow$ **Unbiasedness**
- Moment estimate (a numerical value)

  $\tilde{m}_q = (1/n) \sum_{i=1}^{n} x_i^q$

Some inequalities (notice, underlined quantities denote random variables)
$m_q \neq \underline{\tilde{m}}_q \neq \tilde{m}_q \neq \underline{m}_q$ (three conceptually different mathematical objects)

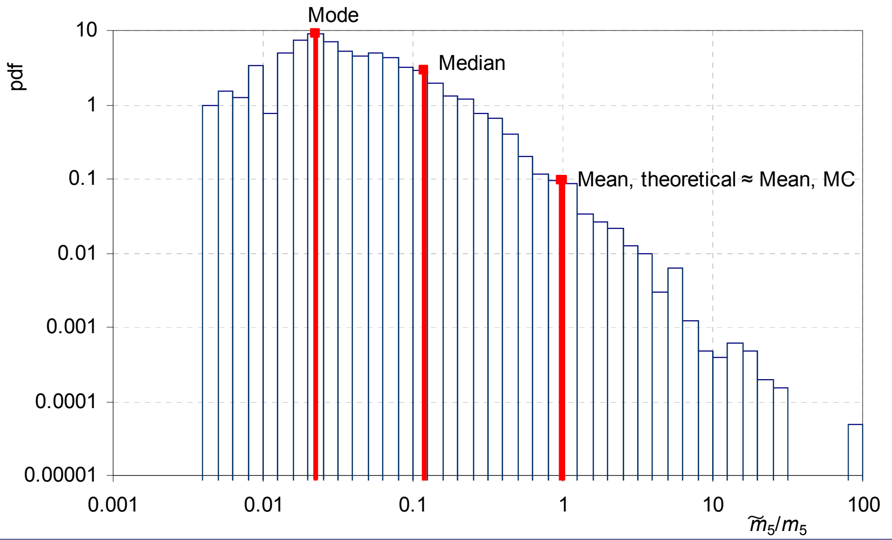
Results of Monte Carlo simulation

- The information content of the empirically estimated moments is high if the distribution of the random variable ($\underline{\tilde{m}}_q / m_q$) is concentrated around 1
- Only low moments (q = 1 and 2) have reasonably low variation
- All others vary within orders of magnitude
- Even the medians are by one or more orders of magnitude lower than 1 for $q > 4$



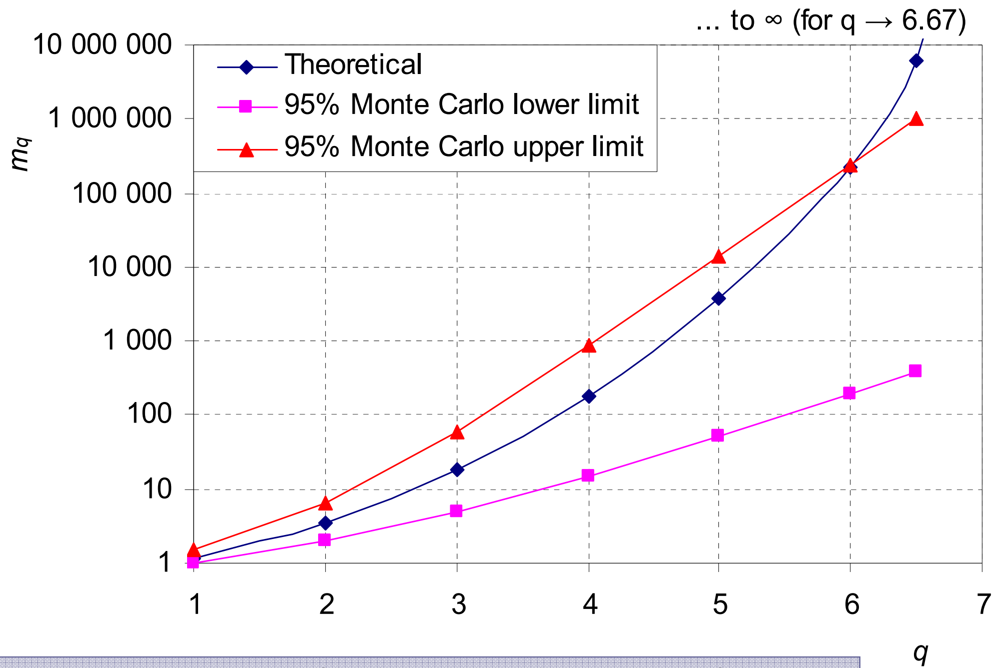
Is there any meaning of theoretical unbiasedness if the probability distribution of the statistical estimator is so broad and skewed?

Results of Monte Carlo simulation: probability density function of $\tilde{m}_5$



Here the bias is theoretically **zero**
However, the probability of calculating (from a unique sample) a value $\tilde{m}_5$ almost **two orders of magnitude less than the true value** (the mode) is **two orders of magnitude higher** than the probability of obtaining the true value (the mean)

Results of Monte Carlo simulation: confidence limits



Even bracketing the true value of high moments between confidence limits may be impossible

Misuse case 2: Bias induced even to 2nd order statistics due to temporal dependence

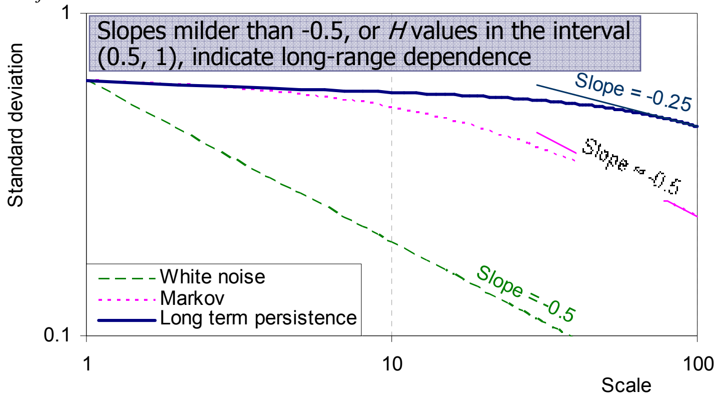
- Dependence is viewed through the autocorrelogram ρ_j (for lag j) of the process or else through the standard deviation $\sigma^{(k)}$ of the time averaged process at scale k:

$$\underline{x}_i^{(k)} := \frac{1}{k} \sum_{l=(i-1)k}^{ik} \underline{x}_l$$

- $\sigma^{(k)}$ is related to ρ_j by a simple transformation, i.e.,

$$\sigma^{(k)} = \frac{\sigma}{\sqrt{k}} \sqrt{\alpha_k} \,, \quad \alpha_k = 1 + 2\sum_{j=1}^{k-1}\left(1 - \frac{j}{k}\right)\rho_j \leftrightarrow \rho_j = \frac{j+1}{2}\alpha_{j+1} - j\alpha_j + \frac{j-1}{2}\alpha_{j-1}$$

- The plot of $\sigma^{(k)}$ vs. k has been termed the climacogram
- The asymptotic slope (high k) in a logarithmic plot is a characteristic of scaling defining the so-called Hurst coefficient: $H = 1 + \text{slope}$



Long-range dependence: The Hurst-Kolmogorov (HK) process

The simplest process with long-range dependence (long-term persistence), the Hurst-Kolmogorov process (after Hurst, 1951; Kolmogorov, 1940; see also Koutsoyiannis and Cohn, 2008), has constant slope of climacogram throughout all scales (power-law climacogram or **perfect time scaling**)

Also its autocorrelogram and power spectrum are power laws of lag j, frequency ω and scale k

Properties of the HK process	At an arbitrary observation scale $k = 1$ (e.g. annual)	At any scale k
Standard deviation	$\sigma \equiv \sigma^{(1)}$	$\sigma^{(k)} = k^{H-1}\sigma$ (can serve as a definition of the HK process; H is the Hurst coefficient; $0.5 < H < 1$)
Autocorrelation function (for lag j)	$\rho_j \equiv \rho_j^{(1)} = \rho_j^{(k)} \approx H(2H-1)\,\lvert j\rvert^{2H-2}$	
Power spectrum (for frequency ω)	$s(\omega) \equiv s^{(1)}(\omega) \approx 4(1-H)\,\sigma^2\,(2\omega)^{1-2H}$	$s^{(k)}(\omega) \approx 4(1-H)\,\sigma^2\,k^{2H-2}\,(2\omega)^{1-2H}$

Short-range dependence: The Markovian process (AR(1))

The simplest process with short-range dependence (short-term persistence), the Markovian process (or the AR(1) process), has autocorrelation defined by a single parameter $\rho \equiv \rho_1$. In this it resembles the HK process. However, in contrast to the HK process, the climacogram does not have a constant slope throughout all scales

Its autocorrelogram is an exponential law and, thus, tends to zero rapidly for increasing lag and/or scale (Koutsoyiannis, 2002)

Properties of the AR(1) process	At scale $k = 1$	At any scale k
Variance	$\gamma_0 \equiv \gamma_0^{(1)}$	$\gamma_0^{(k)} = \gamma_0 \dfrac{k(1-\rho^2) - 2\rho(1-\rho^k)}{k^2(1-\rho)^2}$
Autocorrelation function (for lag j)	$\rho_j = \rho^j$	$\rho_1^{(k)} = \dfrac{\rho(1-\rho^k)^2}{k(1-\rho^2) - 2\rho(1-\rho^k)}$, $\rho_j^{(k)} = \rho_1^{(k)} \rho^{k(j-1)}$
Power spectrum (for frequency ω)	$s_\gamma(\omega) = s_\gamma^{(1)}(\omega)$	$s_\gamma^{(k)}(\omega)/\gamma_0^{(k)} = 2 + 4\rho_1^{(k)} \dfrac{\cos(2\pi\omega) - \rho^k}{1 + \rho^{2k} - 2\rho^k\cos(2\pi\omega)}$

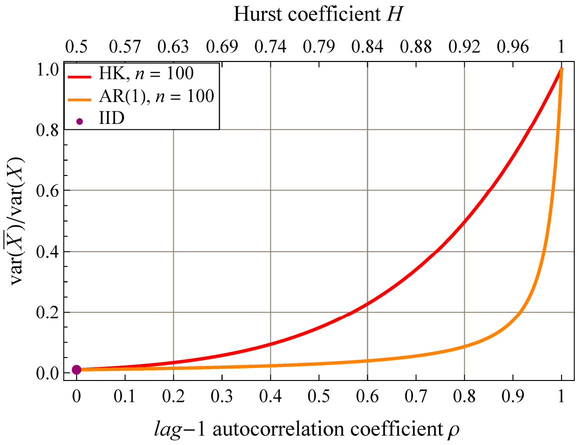
Impacts on statistical estimation: Hurst-Kolmogorov statistics (HKS) vs. classical statistics (CS)

True values →	Mean, μ	Standard deviation, σ	Autocorrelation ρ_l for lag l
Standard estimator	$\underline{\bar{x}} := \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$	$\underline{s} := \sqrt{\frac{1}{n-1}}\sqrt{\sum_{i=1}^{n}(\underline{x}_i - \underline{\bar{x}})^2}$	$\underline{r}_l := \frac{1}{(n-1)\underline{s}^2}\sum_{i=1}^{n-l}(\underline{x}_i - \underline{\bar{x}})(\underline{x}_{i+l} - \underline{\bar{x}})$
Relative bias of estimation, CS	0	≈ 0	≈ 0
Relative bias of estimation, HKS	0	$\approx \sqrt{1 - \frac{1}{n'}} - 1 \approx -\frac{1}{2n'}$	$\approx -\frac{1/\rho_l - 1}{n' - 1}$
Standard deviation of estimator, CS	$\frac{\sigma}{\sqrt{n}}$		
Standard deviation of estimator, HKS	$\frac{\sigma}{\sqrt{n'}}$		

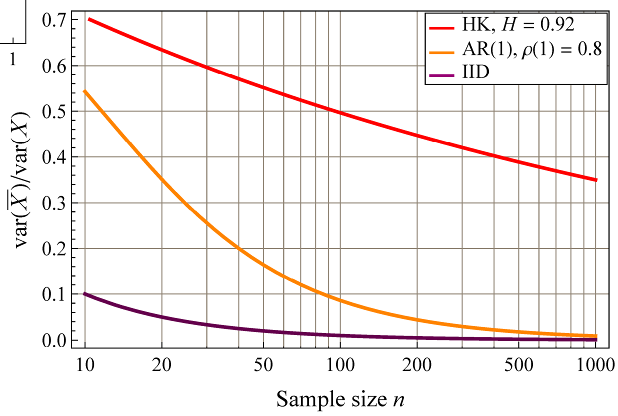
Note: $n' := n^{2-2H}$ is the "equivalent" or "effective" sample size: a sample with size n' in CS results in the same uncertainty of the mean as a sample with size n in HKS (Koutsoyiannis, 2003; Koutsoyiannis & Montanari, 2007).

Note 2: The same relationships hold (approximately) even for Markov processes but with n' defined as $n' := n\dfrac{(1-\rho)^2}{(1-\rho^2) - 2\rho(1-\rho^n)/n}$ (Koutsoyiannis, 2002; Koutsoyiannis & Montanari, 2007).

The variance of the estimator of mean

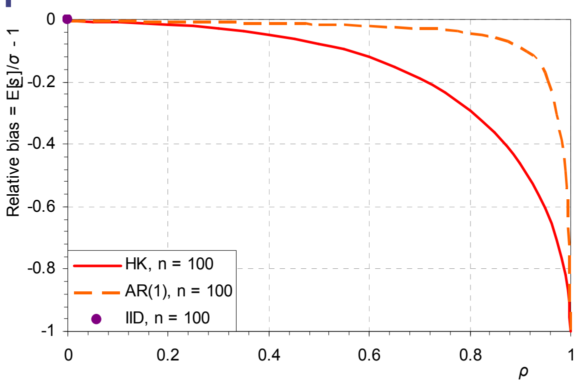


The variance of the estimator of the mean of n IID random variables is σ^2_X/n. Thus, the ratio plotted here for n = 100 is only 0.01. For a Markovian process with, e.g., ρ = 0.8 the ratio is approximately 10 times greater, and for an HK process with H = 0.92 (so that ρ_1 = 0.8) is about 50 times greater!

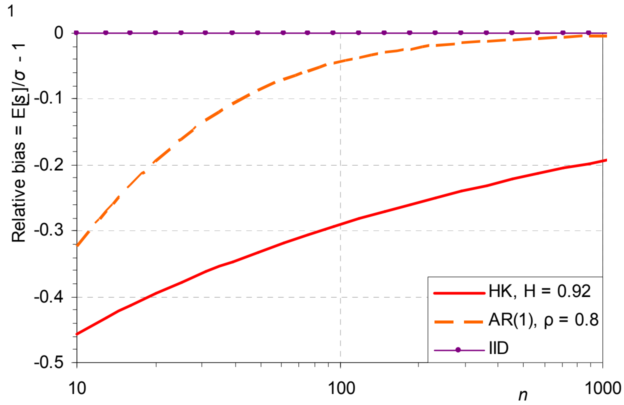


In the same example, while for small samples the Markovian model results in much higher ratio than in the IID case, as the sample size increases the ratio quickly converges to the IID case. In contrast, in the HK case the convergence is extremely slow.

Bias of the classical estimator of standard deviation

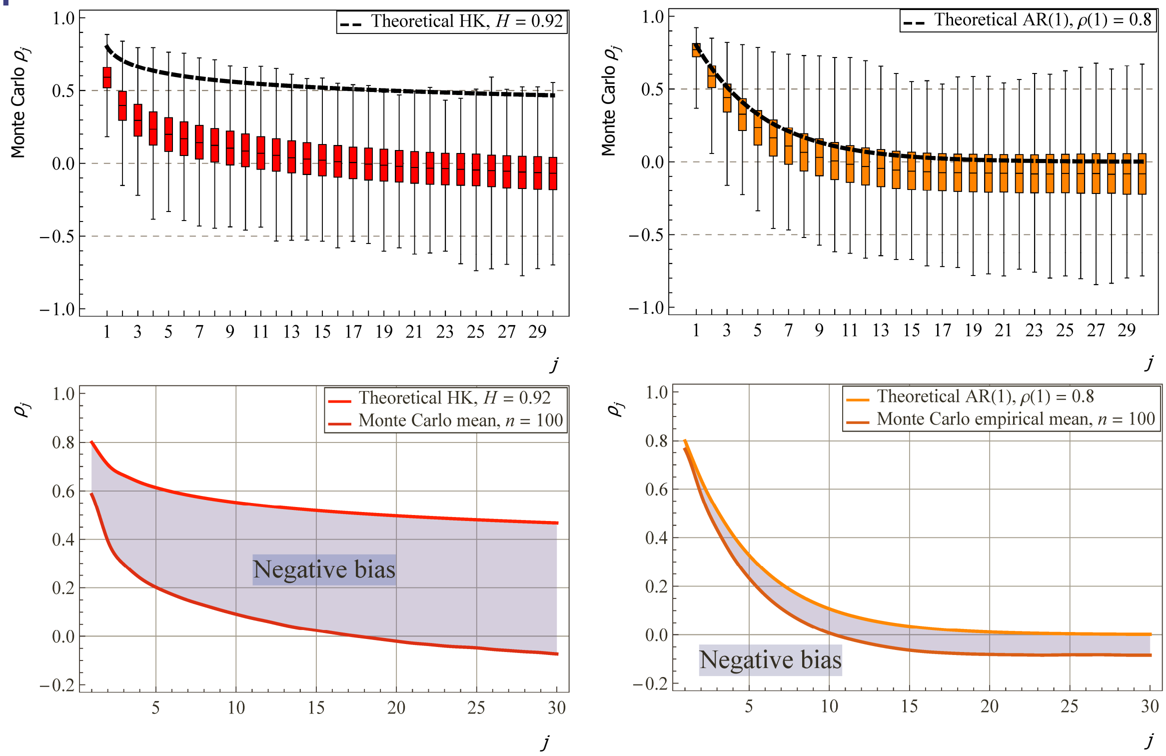


The classical estimator of standard deviation is approximately unbiased for IID random variables. However, as the temporal dependence becomes stronger, the estimator becomes more and more biased. Especially for the HK process with high H, the estimator is severely biased.

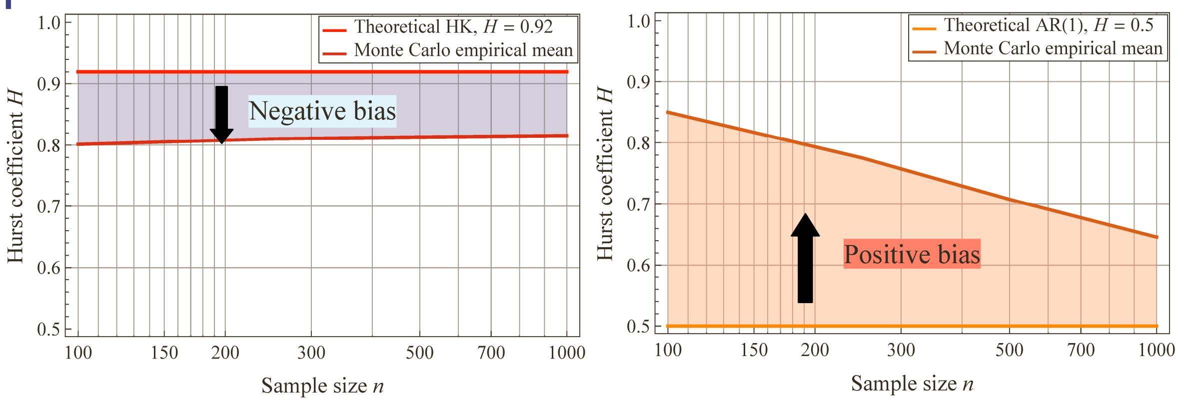


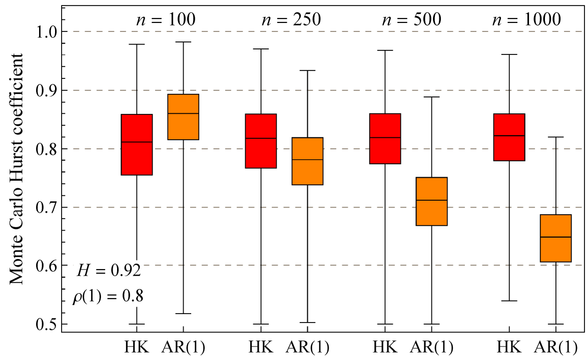
There is a large difference in the bias of the estimator of standard deviation between the Markovian and the HK case. The bias in the HK case for n = 1000 equals that of the Markovian case for n = 20.

Bias of the classical estimator of autocorrelation



Bias in the classical estimation of the Hurst coefficient





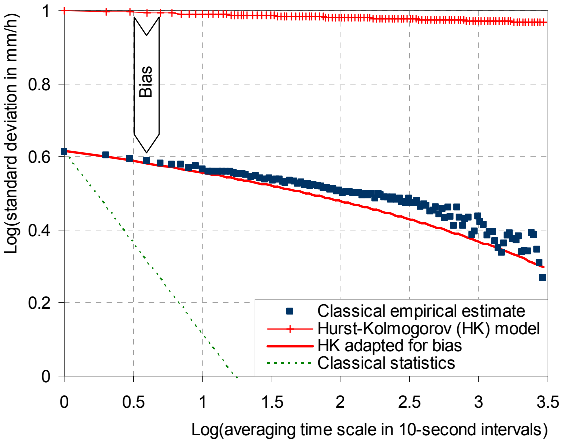
The Hurst coefficients were estimated from the slope of the climacogram using the classical estimator of standard deviation

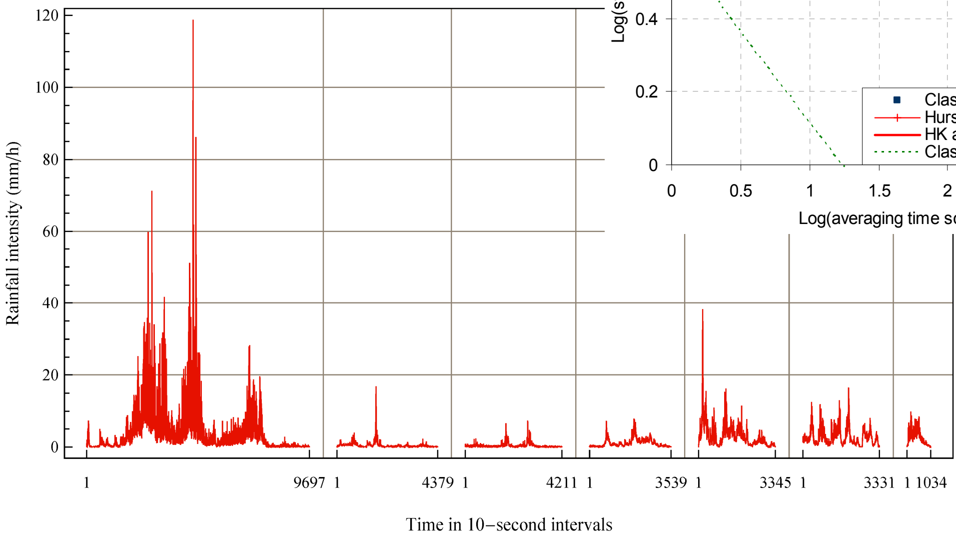
Obviously, this method is inappropriate and demands extremely large samples to estimate the true Hurst coefficient value (for better methods see Tyralis and Koutsoyiannis, 2010)

Example 1: Iowa fine resolution rainfall

Seven storm events of high temporal resolution, recorded by the Hydrometeorology Laboratory at the Iowa University (Georgakakos et al., 1994)

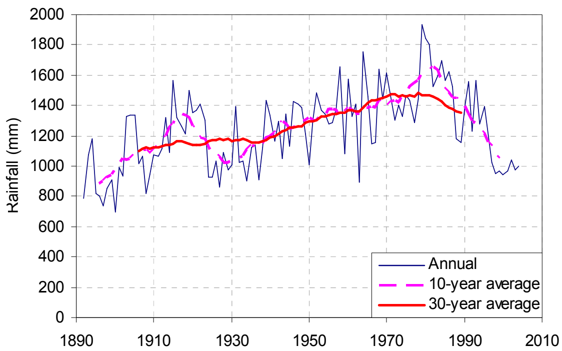
The unified sample suggests an HK behaviour with a very high Hurst coefficient: $H \approx 0.99$

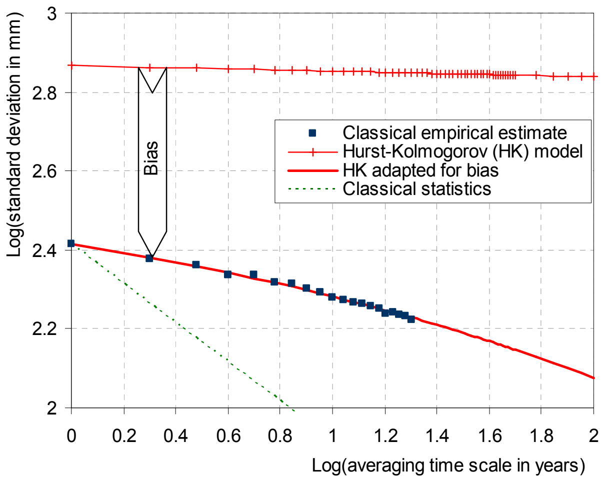




Example 2: The annual rainfall in Maatsuyker Island (Australia)

Maatsuyker Island Lighthouse (Australia), coordinates: -43.65N, 146.27E, 147 m, WMO station code: 94962
Data: 1892-2004, available from http://climexp.knmi.nl/getprcpall.cgi?someone@somewhere+94962+MAATSUYKER_ISLAND_LIGHTHOUSE+

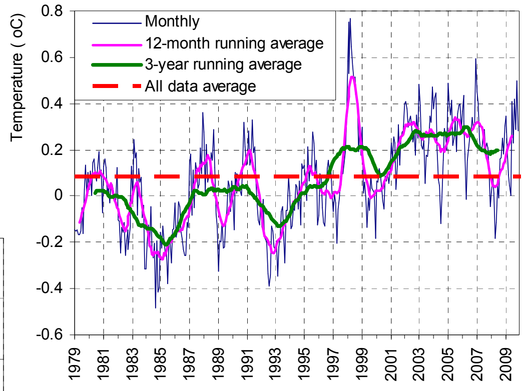


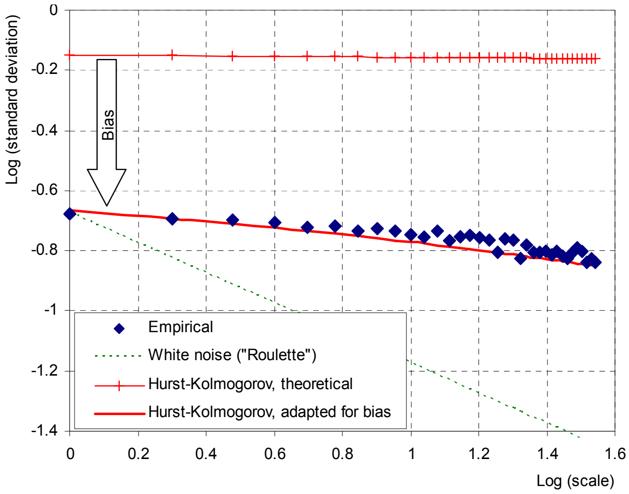


The time series suggests an HK behaviour with a very high Hurst coefficient: $H \approx 0.99$

Example 3: The lower tropospheric temperature

The global average tropospheric temperature estimated from satellite data
Available for the period 1979-2010, from http://vortex.nsstc.uah.edu/public/msu/t2lt/tltglhmam_5.2





The time series suggests an HK behaviour with a very high Hurst coefficient: $H \approx 0.99$

Concluding remarks

- The study of natural processes, including hydrological processes, necessarily relies on concepts and tools of stochastics (probability, statistics, and stochastic processes)—even if sometimes the stochastic character of such concepts is hidden behind complicated algorithms
- The abstract objects of stochastics need to be understood before they can be used in application studies
- Popular computer programs have facilitated calculation of numerical values of such objects
- However, such numerical values may distort, or prevent the formation of, a coherent view of the natural behaviours
- Classical statistics rely on explicit or tacit assumptions, such as independence in time and exponential distribution tails
- Such assumptions are invalidated in natural processes, which suggest scaling in state (power-law distribution tails) and in time (long-range dependence)
- These departures of Nature from classical statistical assumptions imply high biases and notoriously increased uncertainty—and these should be kept in mind when exploring and modelling Nature

References

- Georgakakos, K., A. Carsteanu, P. Sturdevant, and J. Cramer, Observation and Analysis of Midwestern Rain Rates, *Journal of Applied Meteorology*, *33*(12), 1433-1444, 1994
- Hurst, H.E., Long term storage capacities of reservoirs, *Trans. Am. Soc. Civil Engrs.*, 116, 776–808, 1951.
- Kolmogorov, A. N., Wienersche Spiralen und einige andere interessante Kurven in Hilbertschen Raum, *Dokl. Akad. Nauk URSS*, 26, 115–118, 1940
- Koutsoyiannis, D., The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrological Sciences Journal*, 47 (4), 573–595, 2002
- Koutsoyiannis, D., Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrological Sciences Journal*, 48 (1), 3–24, 2003
- Koutsoyiannis, D., Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records, *Hydrological Sciences Journal*, 49 (4), 591–610, 2004
- Koutsoyiannis, D., Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling, *Hydrological Sciences Journal*, 50 (3), 381–404, 2005
- Koutsoyiannis, D., Some problems in inference from time series of geophysical processes (solicited), *European Geosciences Union General Assembly 2010, Geophysical Research Abstracts, Vol. 12*, Vienna, EGU2010-14229, European Geosciences Union, 2010 (http://www.itia.ntua.gr/en/docinfo/973/)
- Koutsoyiannis, D., and T.A. Cohn, The Hurst phenomenon and climate (solicited), *European Geosciences Union General Assembly 2008, Geophysical Research Abstracts, Vol. 10*, Vienna, 11804, European Geosciences Union, 2008 (http://www.itia.ntua.gr/en/docinfo/849/)
- Koutsoyiannis, D., and A. Montanari, Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resources Research*, 43 (5), W05429, doi:10.1029/2006WR005592, 2007
- Papalexiou, S.-M., and D. Koutsoyiannis, On the tail of the daily rainfall probability distribution: Exponential-type, power-type or something else?, *European Geosciences Union General Assembly 2009, Geophysical Research Abstracts, Vol. 12*, Vienna, EGU2010-111769, European Geosciences Union, 2010 (http://www.itia.ntua.gr/en/docinfo/977/)
- Tyralis, H., and D. Koutsoyiannis, Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process, *Stochastic Environmental Research & Risk Assessment*, 2010 (in press)