European Geosciences Union
General Assembly 2012

Vienna, Austria, 22 - 27 April 2012

Session: HS7.5/NP8.3
Hydroclimatic stochastics

ON STATISTICAL BIASES AND THEIR COMMON NEGLECT

Available online:
http://itia.ntua.gr/en/docinfo/1205/
Contact Info: smp@itia.ntua.gr

E. Houdalaki, M. Basta, N. Boboti, N. Bountas, E. Dodoula, T. Iliopoulou, S. Ioannidou, K. Kassas, S. Nerantzaki, E. Papatriantafyllou, K. Tettas, D. Tsirantonaki, S.M. Papalexiou, and D. Koutsoyiannis

Department of Water Resources and Environmental Engineering, National Technical University of Athens, Greece (www.itia.ntua.gr)

1. Abstract

The study of natural phenomena such as hydroclimatic processes demands the use of stochastic tools and the good understanding thereof. However, common statistical practices are often based on classical statistics, which assumes independent identically distributed variables with Gaussian distributions. However, in most cases geophysical processes exhibit temporal dependence and even long term persistence. Also, some statistical estimators for nonnegative random variables have distributions radically different from Gaussian. We demonstrate the impact of neglecting dependence and non-normality in parameter estimators and how this can result in misleading conclusions and futile predictions. To accomplish that, we use synthetic examples derived by Monte Carlo techniques and we also provide a number of examples of misuse.

Acknowledgment: This research is conducted within the frame of the undergraduate course "Stochastic Methods in Water Resources" of the National Technical University of Athens (NTUA). The School of Civil Engineering of NTUA provided financial support for the participation of the students in the Assembly.

2. Motivation

- In the last twenty years there has been a large raise of interest in multifractal analyses especially in the study of hydrological processes, particularly in rainfall modelling.
- In such analyses, high order moments are estimated and used in model identification and fitting as if they were reliable.
- Using simple Monte Carlo simulations we find that the reliability of such methods is questionable.
- At the same time in many studies it has been a common practice to neglect the bias in statistical estimations; this bias is introduced when the process exhibits dependence in time and is amplified when the distribution function is non Gaussian.
- Even in quantities that in theory are unbiased, the dependence and non-normality affect significantly their statistical properties and result in huge departures from classical statistical results.
- Therefore we try to indicate the impact of these misconceptions and neglects.

3. Examples from literature

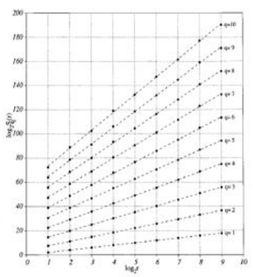


Figure 8. The first 10 structure functions $S_q(r)$ (circles) estimated on one synthetic field, chosen between the 64 generations with the two-dimensional model, are plotted as a function of r. The scaling of structure functions is highlighted by the log-log least squares regressions (dashed curves).

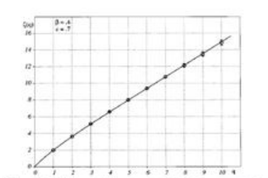


Figure 9. The theoretical expectation (58) of moments $\zeta(q)$ for the two-dimensional ($d = 2$) model with log-Poisson generator with parameters $\beta = 0.4$ and $c = 0.7$ (solid curve). Error bars represent averages and standard deviations of $\zeta(q)$ estimated from the 64 synthetic signals at 1024×1024 resolution.

A brief search in Google resulted in a number of papers in which high order moments are used as if they were reliable. (References not given on purpose).

4. Methodology

We generate 10000 time series with a sample size of 100 values each, using the following models:

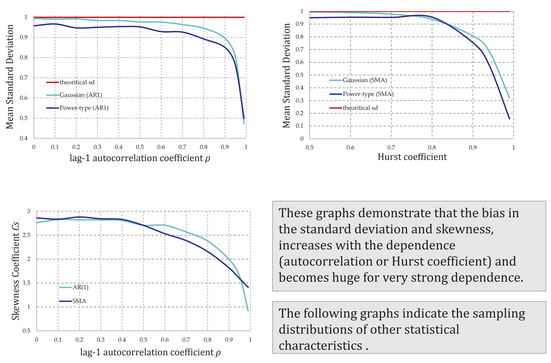
- Autoregressive of order 1 (AR(1)), with lag-one autocorrelation $\rho(1)$: 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99.
- Hurst-Kolmogorov (aka FGN), with Hurst coefficient H: 0.50, 0.57, 0.63, 0.69, 0.74, 0.79, 0.84, 0.88, 0.92, 0.96, 0.98, 0.996 which correspond to the same autocorrelation coefficients as above (for comparison).

The same procedure is followed both for Gaussian N(0,1) and a modified distribution resulting from the Gaussian, using the transformation: $g(z) = \beta\left(\frac{\exp\left(z^2\right)^{\frac{\gamma_1\gamma_2}{2}} - 1}{\gamma_2}\right)^{\frac{1}{\gamma_1}}$

The distribution produced appears to be a power-type distribution with the form:

$$f(x) = \frac{\sqrt{\frac{2}{\pi}}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\left[\gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}+1\right]^{\frac{1}{\gamma_1\gamma_2}-1}}{\beta\sqrt{\frac{2}{\gamma_1\gamma_2}\ln\left[1+\gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right]}}$$ and an asymptotic behavior of tail $x^{-\frac{1}{\gamma_2}} = x^{-4}$

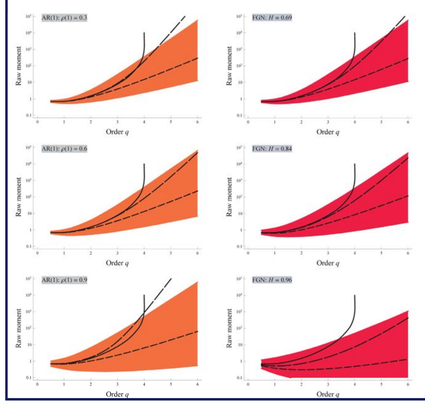
5. Bias in the SD and skewness



These graphs demonstrate that the bias in the standard deviation and skewness, increases with the dependence (autocorrelation or Hurst coefficient) and becomes huge for very strong dependence.

The following graphs indicate the sampling distributions of other statistical characteristics .

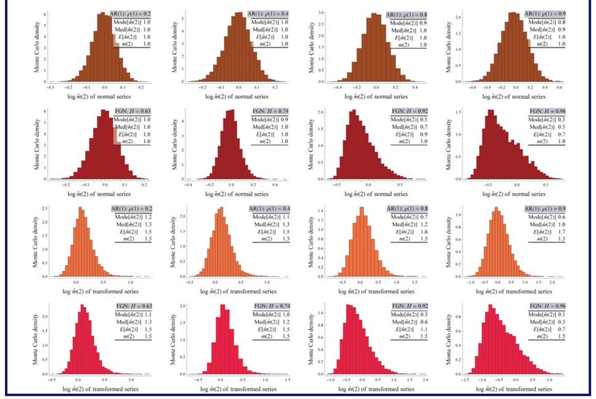
6. Moments confidence interval



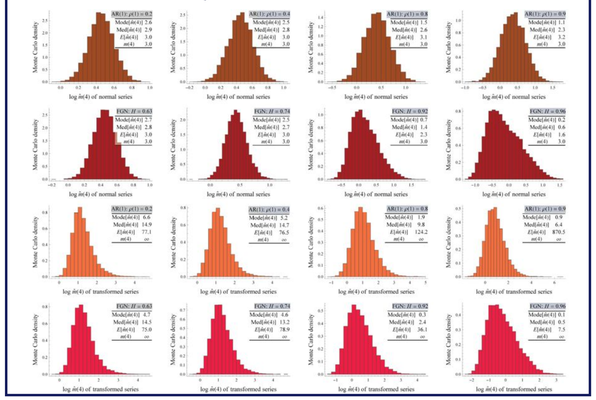
The solid line indicates the theoretical mean of the raw moments, which tends to infinity in the 4th order. Even the empirical mean of raw moments (dashed line with the wide dashes) surpasses the confidence interval for orders higher than 4, something that makes the use of high moments unreliable. The last line (short dashes) indicates the median.

Notice the logarithmic scale on the vertical axis to understand the real size of the confidence interval!

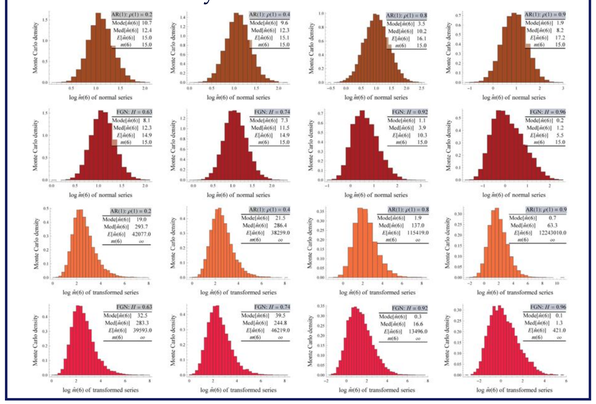
7. Statistical analysis of 2nd moment



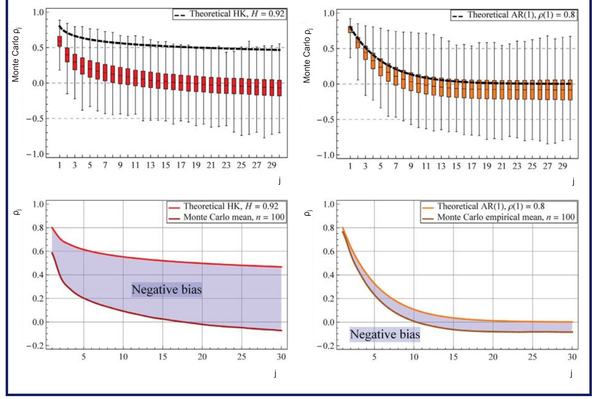
8. Statistical analysis of 4th moment



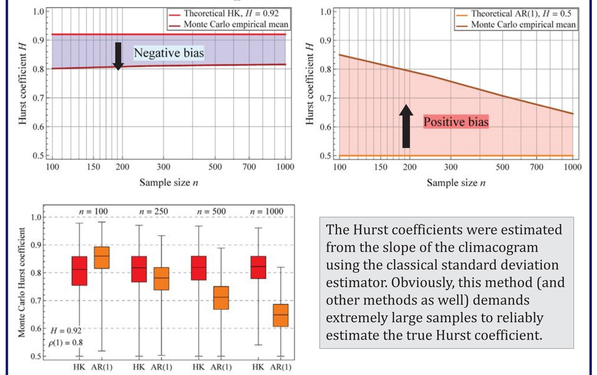
9. Statistical analysis of 6th moment



10. Bias in correlation



11. Bias in Hurst exponent



The Hurst coefficients were estimated from the slope of the climacogram using the classical standard deviation estimator. Obviously, this method (and other methods as well) demands extremely large samples to reliably estimate the true Hurst coefficient.

12. Conclusions

- The study and modelling of natural phenomena, including hydrological processes presupposes the understanding of their stochastic behaviour and, therefore, the correct use of stochastic tools.
- In literature there are many cases of misuse concerning the neglect of dependence and non normality in distributions as well as intense use of estimates of high order moments as if they were reliable.
- By applying Monte-Carlo simulation for both normal and non normal distributions we demonstrate that :
  - ✓ In cases of dependence there is significant negative bias in estimating standard deviation, skewness, autocorrelation and Hurst coefficient which may lead in false conclusions.
  - ✓ The calculation of numerical values of high order moments is misleading as the theoretical moments may tend to infinity for high orders, while the sample estimates are always finite. Even smaller order moments can be very uncertain.
- Finally, the use of high moments provides no reliable or even meaningful result. In addition, the possible bias on the statistical properties should be taken into account.