IUGG XXIV General Assembly 2007
Perugia, Italy, 2 - 13 July 2007

International Association of Hydrological Sciences,
Session HW2003 - Analysis of Variability in Hydrological Data Series

Stochastic modelling of skewed data exhibiting long-range dependence

S.M. Papalexiou

Department of Water Resources, National Technical University of Athens

1. Abstract

Time series with long-range dependence appear in many fields including hydrology and there are several studies that have provided evidence of long autocorrelation tails. Provided that the intensity of the long-range dependence in time series of a certain process, quantified by the self-similarity parameter, also known as the Hurst exponent H, could not be falsified, it is then essential that the variable of interest is modelled by a model reproducing long-range dependence. Common models of this category that have been widely used are the fractional Gaussian noise (FGN) and the fractional ARIMA (FARIMA). In case of a variable exhibiting skewness, the previous models can not be implemented in a direct manner. In order to preserve skewness in the simulated series, a normalizing transformation is typically applied in the real-life data at first. The models are then fitted to the normalized data and the produced synthetic series are finally de-normalized. In this paper, a different method is proposed, consisting of two parts. The first one regards the approximation of the long-range dependence by an autoregressive model of high order p AR(p), while the second one regards the direct calculation of the main statistical properties of the random component, that is mean, variance and skewness coefficient. The skewness coefficient calculation of the random component is done using joint sample moments. The advantage of the method is its efficiency and simplicity and the analytical solution.

2. Motivation

- Since Hurst (1951) observed the long-term persistence phenomenon in the annual average streamflows of Nile, the same behaviour has been identified in numerous natural processes while, its importance has been underlined by scientists in many controversial disciplines. It seems that the Hurst phenomenon is ubiquitous in nature and this makes it necessary to find adequate ways to model it.
- Many models have been proposed in the literature that preserve the Hurst behaviour, such as Fractional Gaussian Noise (FGN) (i.e. Mandelbrot, 1969; Mandelbrot and Wallis 1969), fast FGN (Mandelbrot, 1971), broken line models (i.e. Ditlevsen, 1971), fractional ARIMA (Hosking, 1981), and recently symmetric moving average models (SMA) (Koutsoyiannis, 2000; 2002).
- If the Hurst behaviour appears in a process, it needs to modeled as it affects dramatically the time series structure. Another distinguished characteristic of hydrological processes, that needs to be modeled, is asymmetry. In this direction have been made many attempts to adapt standard models to preserve the skewness (i.e. Matalas and Wallis, 1976).
- Some of the previous models are not easy to apply as the parameters are not easy to estimate. while other can preserve the skewness but not the Hurst behaviour and vice versa. Other problems are the narrow type of autocorrelation functions that those model can simulate (exception is the SMA model).
- In this study is proposed a general methodology to preserve both the Hurst behaviour and skewness. The framework of the methodology is simple: the Hurst phenomenon is modeled from an autoregressive model of high order, AR(p), while the skewness is preserved by evaluating the skewness coefficient of the random component of the model. The model should be easy to apply and suitable for any practical purposes such as hydrologic design or water resources management.

3. Modelling Approach

- In order to preserve the long-range dependence or the Hurst phenomenon in the simulated time series, a high order autoregressive model is implemented. The long-range dependence behaviour, is essentially the slow decay of the autocorrelation function with time. On the contrary, the AR(*p*) models are considered to be short-range dependence models. Nevertheless, as this study reveals, AR(*p*) models of high order can reproduce the Hurst phenomenon sufficiently enough for any practical modelling purposes.
- In the general case of order *p*, the AR(*p*) model takes the following form: $X_t = \varepsilon_t + \sum_{i=1}^{p} X_{t-i}\, \alpha_i$

  where ε_t is the innovation or the random component and a_i are coefficients. In order to fit the model to a dataset, the a_i coefficients and the basic statistics (mean, standard deviation) of the ε_t have to be estimated.
- The auto-covariance function γ_k of the AR(*p*) model for lag k and for k>0 is given by

$$\gamma_k = \sum_{i=1}^{p} \alpha_i\, \gamma_{|i-k|}$$

  The replacement of γ_k with the samples estimates and the implementation of the last equation *p* times gives a linear system of equations that can be solved straightforwardly, evaluating therefore the a_i coefficients.
- Finally, the mean and the variance of the ε_t can be estimated using the following two equations.

$$\mu_{\varepsilon t} = \mu_{X_t}\left(1 - \sum_{i=1}^{p} \alpha_i\, \gamma_i\right) \qquad \sigma_{\varepsilon t}^2 = \gamma_0 - \sum_{i=1}^{p} \alpha_i\, \gamma_i$$

4. Preserving the Skewness in an AR(p) Model

- To preserve asymmetry in the simulated time series, it is necessary to evaluate the skewness coefficient of the innovation, Csk_{ε}. It can be shown that the third central moment of the innovation of the AR(p) model is

$$\mu 3_{\varepsilon_t} = \mu 3_{X_t} - \mathrm{E}\left(\sum_{i=1}^{p} a_i X_{t-i}\right)^3 \qquad (1)$$

- Defining as multi-auto-covariance of order (m_1, m_2,…,m_n) and lag (l_1, l_2,…, l_3),

$$\mu_{(m_1,m_2\ldots m_n)}^{(l_1,l_2\ldots l_n)} = E\left(\left(X_{t-l_1} - \mu_X\right)^{m_1}\left(X_{t-l_2} - \mu_X\right)^{m_2}\ldots\left(X_{t-l_3} - \mu_X\right)^{m_n}\right)$$

it can be proven that the following equation is valid,

$$\mathrm{E}\left(\sum_{i=1}^{p} a_i X_{t-i}\right)^3 = \sum_{i=1}^{p}\mu_{(3)}^{(0)} a_i^3 + 3\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\mu_{(2,1)}^{(0,j-i)} a_i^2 a_j + 3\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\mu_{(1,2)}^{(0,j-i)} a_i a_j^2 + 6\sum_{i=1}^{p-2}\sum_{j=i+1}^{p-1}\sum_{k=j+1}^{p}\mu_{(1,1,1)}^{(0,j-i,k-i)} a_i a_j a_k$$

- Replacing the multi-auto-covariance terms in the previous equation with the sample estimates, given by

$$\widehat{\mu}_{(m_1,m_2\ldots m_n)}^{(l_1,l_2\ldots l_n)} = \frac{1}{k-\max\left(l_1,l_2\ldots l_n\right)+1}\sum_{i=1}^{k-\max(l_1,l_2\ldots l_n)+1}\left(\left(x_{i+l_1} - \widehat{\mu}_X\right)^{m_1}\left(x_{i+l_2} - \widehat{\mu}_X\right)^{m_2}\ldots\left(x_{i+l_3} - \widehat{\mu}_X\right)^{m_n}\right)$$

it is then straightforward to estimate the $\mu_{3\varepsilon}$ in (1) and thus the $\widehat{C}_{sk_\varepsilon} = \dfrac{\widehat{\mu}_{3_\varepsilon}}{\widehat{\sigma}_\varepsilon^3}$

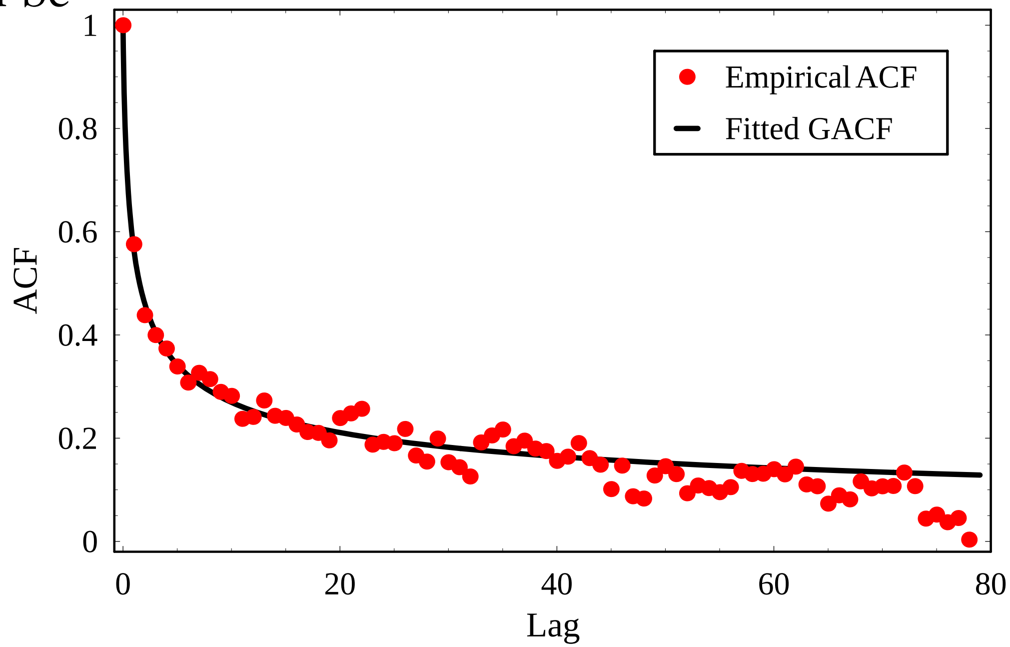
5. The Generalized AutoCorrelation Function (GACF)

- The major criticism of a high order AR(*p*) model would focus on the lack of parsimony, as estimation of the autocorrelation function up to lag *p* is required to fit the model. Moreover, it is well known that the estimator of the ACF is highly variable and that it increases its variability with increasing lag (Bras and Rodriquez, 1985). Consequently, the uncertainty in the estimation of the ACF would lead to uncertain validation of the model parameters. To overcome this disadvantage, it is proposed to fit a generalized ACF, $\rho^{(G)}$,

$$\rho_j^{(G)} = \left(\alpha\,\beta\, j^{\delta} + 1\right)^{-\frac{1}{\beta}}$$

to the first few empirical ACF values (where a, β, δ are positive parameters and j is the lag). Subsequently, the fitted GACF can be used to extrapolate ACF values for high j.

The figure depicts the empirical autocorrelation function of the Nilometer dataset (analysis follows) and the fitted GACF. The GACF has been fitted to the first 10 empirical values of the ACF by minimizing the square error.



6. The Generalized Lambda Distribution (GLD)

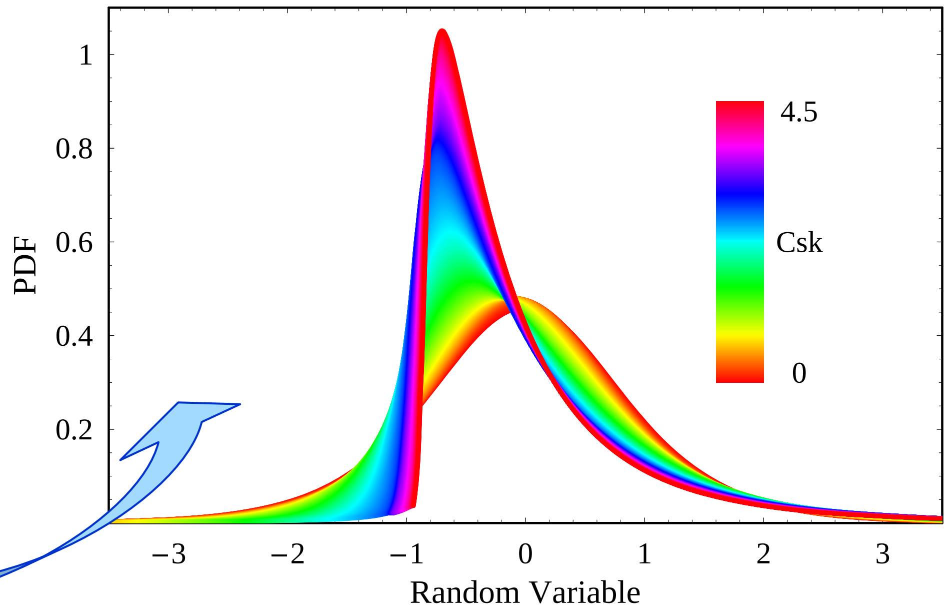
- In order to preserve the skewness in the simulated series, the innovation ε_t must be sampled from a distribution with variable skewness . Such a flexible distribution is the GLD family.
- The GLD(λ_1, λ_2, λ_3, λ_4) family distributions originated from the one-parameter lambda distribution proposed by John Tukey (1960) and was generalized for Monte Carlo simulation purposes by John Ramberg and Bruce Schmeiser (1974). Although the GLD has been applied in many fields since the early 1970s (Karian and Dudewicz, 2000), it has never been used in hydrology.
- The GLD family with parameters λ_1, λ_2, λ_3 and λ_4, is defined in terms of its percentile function,

$$Q(y) = \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2} + \lambda_1$$

where $0 < y < 1$. The parameters λ_1 and λ_2 are, respectively, location and scale parameters, while λ_3 and λ_4 determine the skewness and kurtosis of the distribution.
The GLD probability density function is

$$f(x) = \frac{\lambda_2}{\lambda_3\, y^{\lambda_3 - 1} + (1-y)^{\lambda_4 - 1}\, \lambda_4} \quad \text{at } x = Q(y)$$

- The restrictions on λ_1, λ_2, λ_3 and λ_4, that yield a valid GLD distribution, the parameter space and the skewness–kurstosis space are discussed in detail by Karian and Dudewicz (2000). In the next figure GLD pdfs are plotted with mean = 0, variance = 1 and skewness coefficient *C*sk ranging from 0 – 4.5



7. Fitting the GLD and Sampling

- If X is GLD(λ_1, λ_2, λ_3, λ_4) with λ_3>-1/4 and λ_4>-1/4, then its first four moments (Ramberg et al., 1979), μ, μ_2, μ_3, μ_4 (mean, variance, skewness coefficient, and kurtosis coefficient), are given by

$$\mu = \frac{A}{\lambda_2} + \lambda_1 \qquad \mu_2 = \frac{B - A^2}{\lambda_2^2} \qquad C_{sk} = \frac{\mu_3}{\sigma^3} = \frac{2A^3 - 3BA + C}{\sigma^3 \lambda_2^3} \qquad C_k = \frac{\mu_4}{\sigma^4} = \frac{-3A^4 + 6BA^2 - 4CA + D}{\sigma^4 \lambda_2^4}$$

where $A = \frac{1}{\lambda_3 + 1} - \frac{1}{\lambda_4 + 1}$

$$B = -2\,\mathrm{B}(\lambda_3 + 1, \lambda_4 + 1) + \frac{1}{2\lambda_3 + 1} + \frac{1}{2\lambda_4 + 1}$$

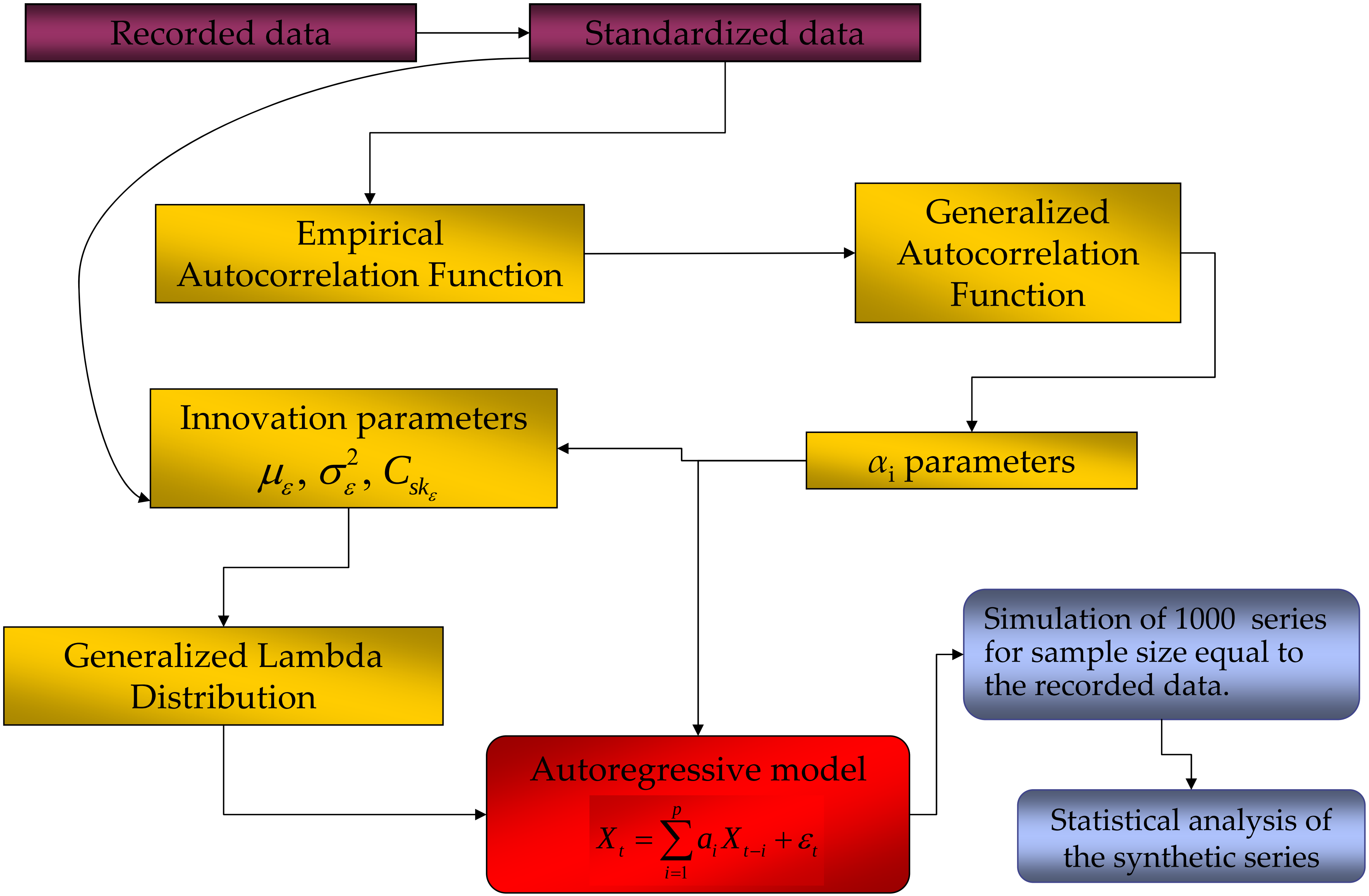
$$C = 3\,\mathrm{B}(\lambda_3 + 1, 2\lambda_4 + 1) - 3\,\mathrm{B}(2\lambda_3 + 1, \lambda_4 + 1) + \frac{1}{3\lambda_3 + 1} - \frac{1}{3\lambda_4 + 1}$$

$$D = -4\,\mathrm{B}(\lambda_3 + 1, 3\lambda_4 + 1) + 6\,\mathrm{B}(2\lambda_3 + 1, 2\lambda_4 + 1) - 4\,\mathrm{B}(3\lambda_3 + 1, \lambda_4 + 1) + \frac{1}{4\lambda_3 + 1} + \frac{1}{4\lambda_4 + 1}$$

and B is the Beta function defined as $\mathrm{B}(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1}\, dt$

- If we consider the innovation ε_t as a random variable with known estimation of the mean, the variance, the skewness and the kurtosis coefficient, a GLD distribution can be fitted by solving numerically the previous nonlinear system. The mean, the variance and the skewness coefficient of ε_t can be analytically estimated as described in slide four. At the moment there is no analytical way to estimate the kurtosis coefficient of ε_t, but heuristically for this study was taken the minimum so as λ_3<0 and λ_4<0, which implies that the fitted GLD ranges form -∞ to ∞ (Karian and Dudewicz, 2000).
- Once the parameters λ_1, λ_2, λ_3, and λ_4 of the GLD are estimated, the sampling is very easy as the percentile function has a simple and analytical formulae.

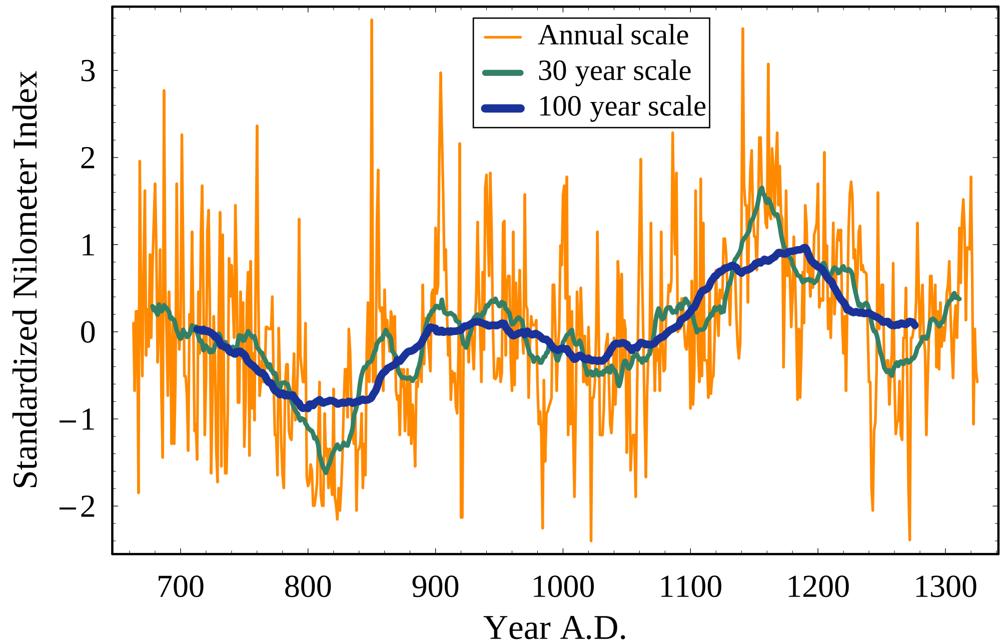
8. Simulation Organogram



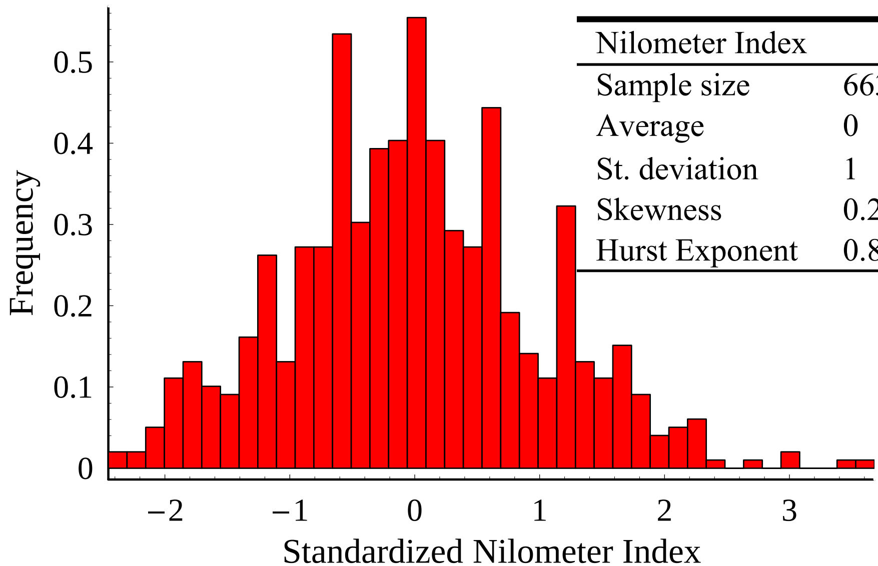
9. Original Data I: Nilometer Index

Standardized Nilometer series indicating the annual minimum water level of the Nile river for the years 622 to 1284 A.D.

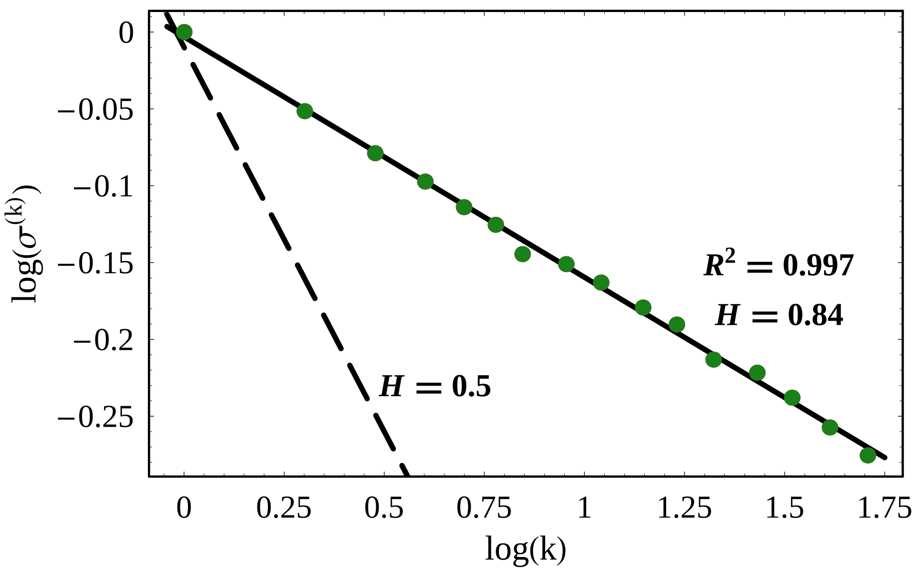
(663 years; Beran, 1994)



A data with small positive skewness but with a large Hurst exponent value that verifies the multiscale fluctuations.



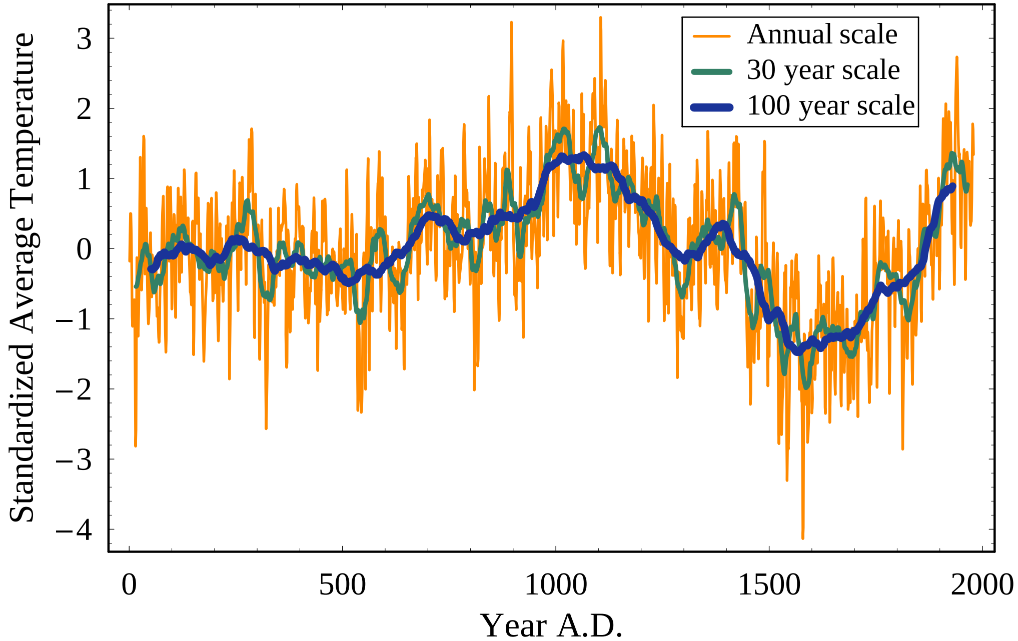
Nilometer Index	
Sample size	663
Average	0
St. deviation	1
Skewness	0.24
Hurst Exponent	0.84



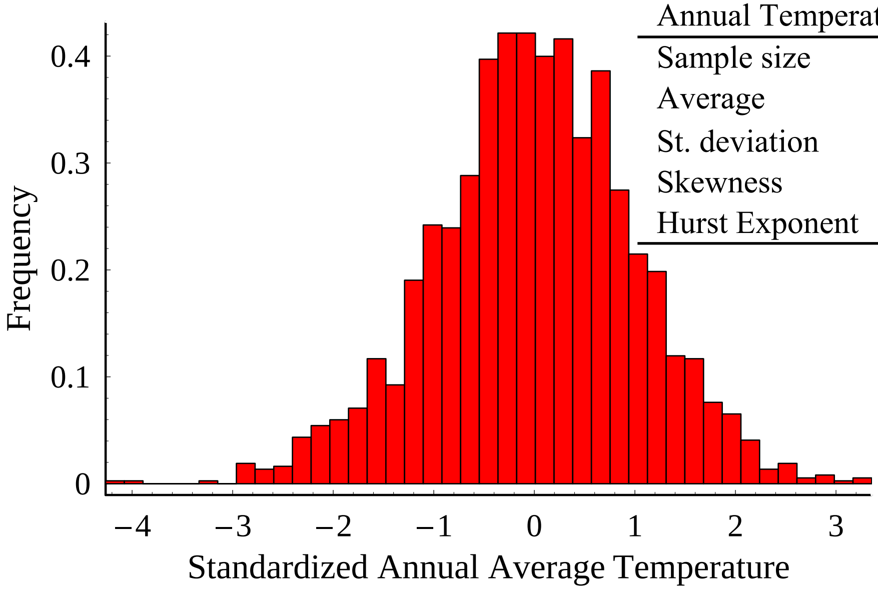
10. Original Data II: Annual Temperatures

Northern Hemisphere temperature anomalies in °C with reference to 1961–1990 mean (Standardized in the figure on the right).

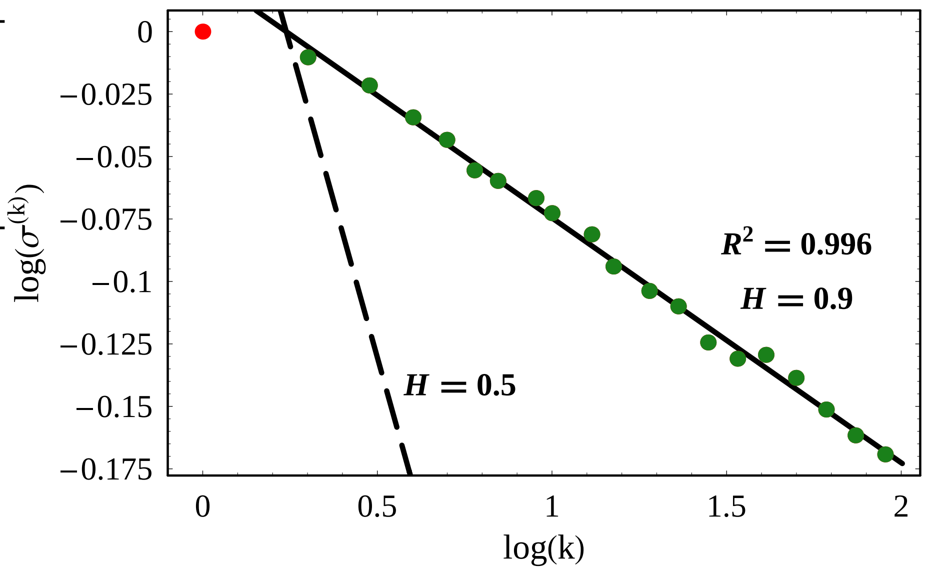
(2000 years, Moberg et al., 2005)



A highly variable Northern Hemisphere temperature reconstruction that reveals the large natural variability of the climate in multiple scales. The multiscale variation is verified by the large Hurst exponent value.



Annual Temperatures	
Sample size	1979
Average	0
St. deviation	1
Skewness	-0.13
Hurst Exponent	0.9

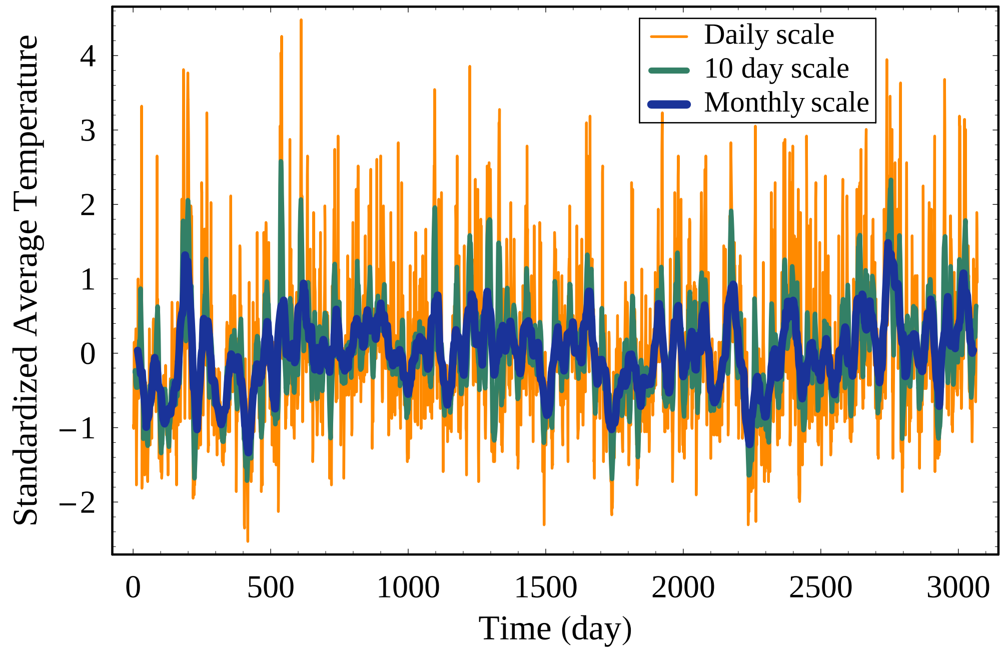


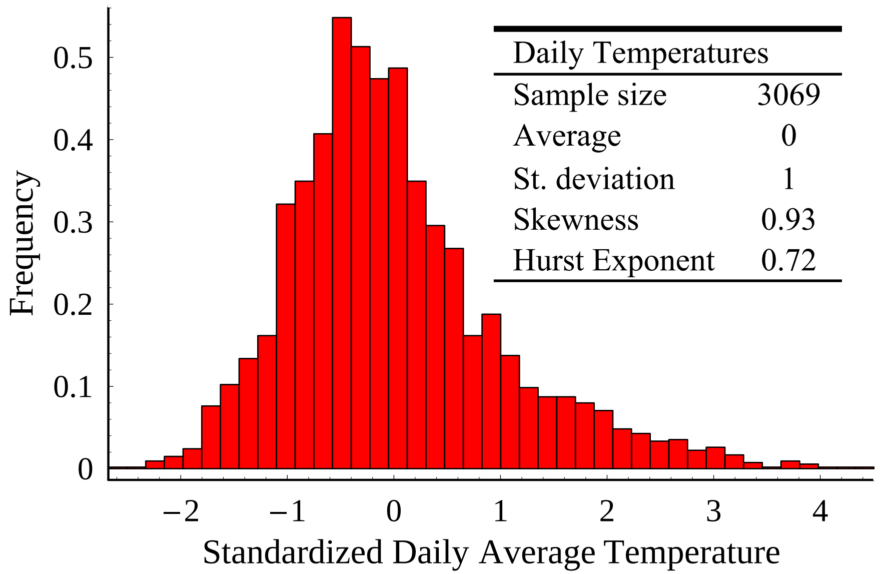
11. Original Data III: Daily Average Temperatures

Standardized average daily temperatures in July recorded at Den Helder station in Netherlands, from 1901 to2005.

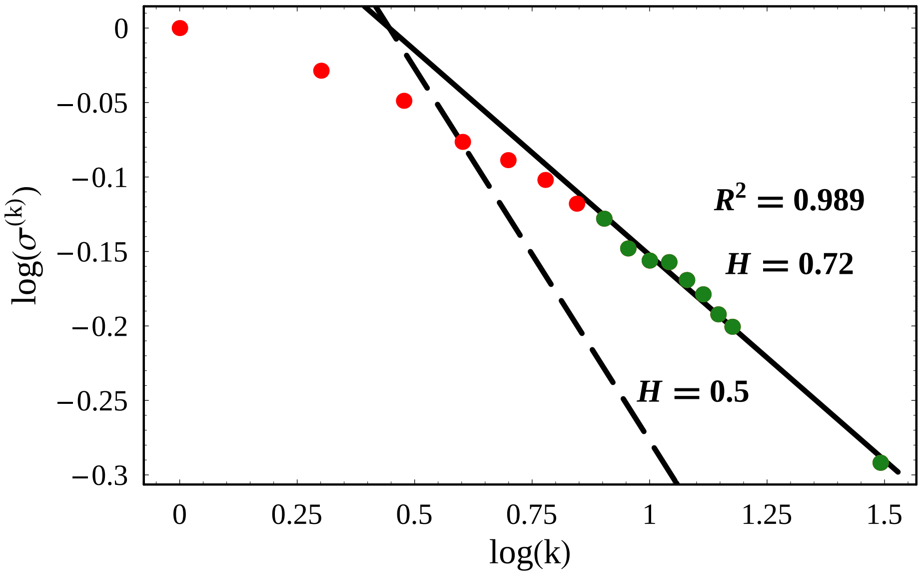
(source: Royal Netherlands Meteorological Institute)

The histogram below depicts the positive asymmetry of the dataset, while the long-range dependence is manifested from the high Hurst exponent value.





Daily Temperatures	
Sample size	3069
Average	0
St. deviation	1
Skewness	0.93
Hurst Exponent	0.72

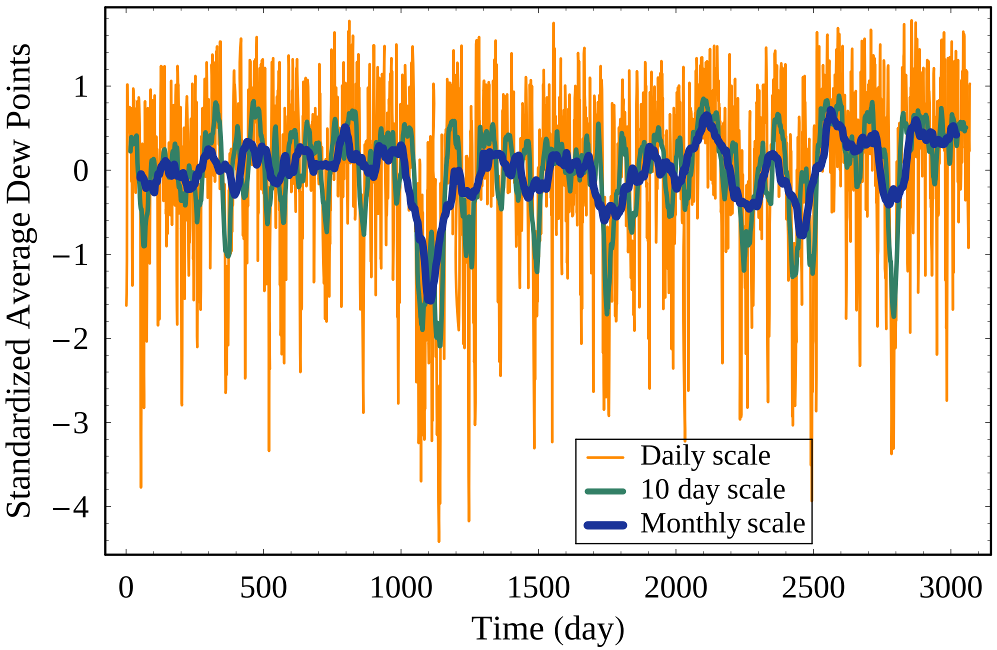


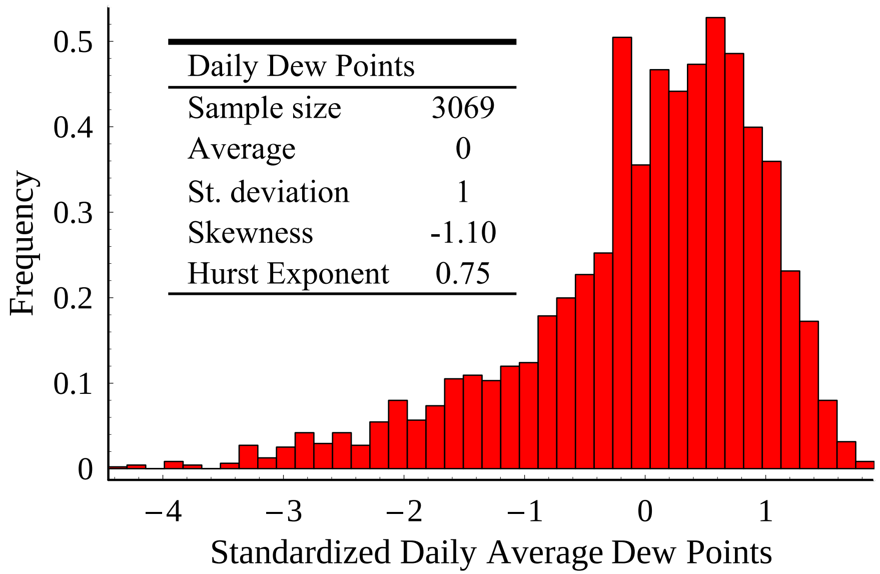
12. Original Data IV: Daily Average Dew Points

Standardized average daily dew points in January recorded at Den Helder station in Netherlands, from 1901 to 2005.

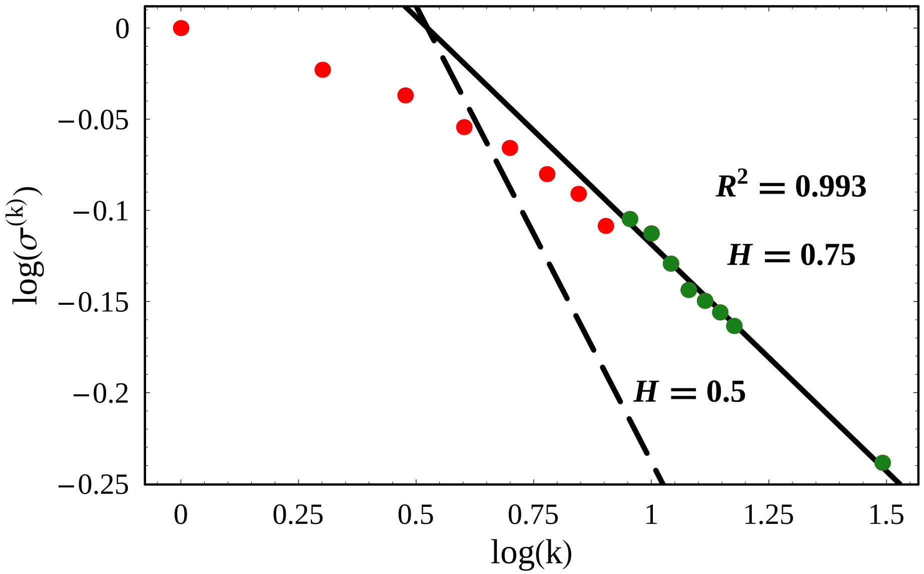
(source: Royal Netherlands Meteorological Institute)

This dataset is suitable for the purposes of this study as it shows high negative asymmetry and a high Hurst exponent value.



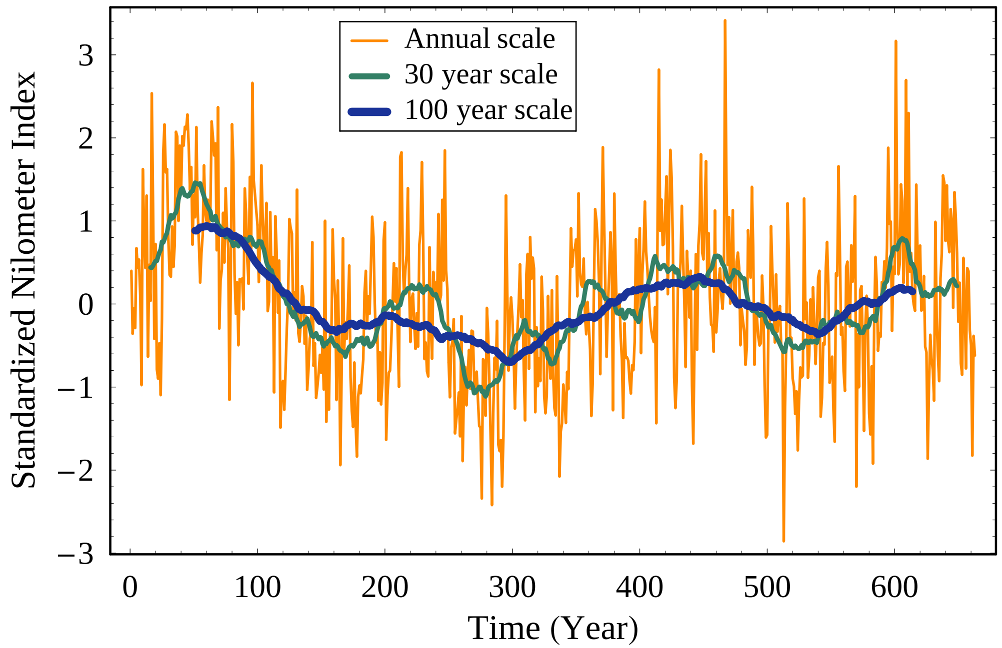


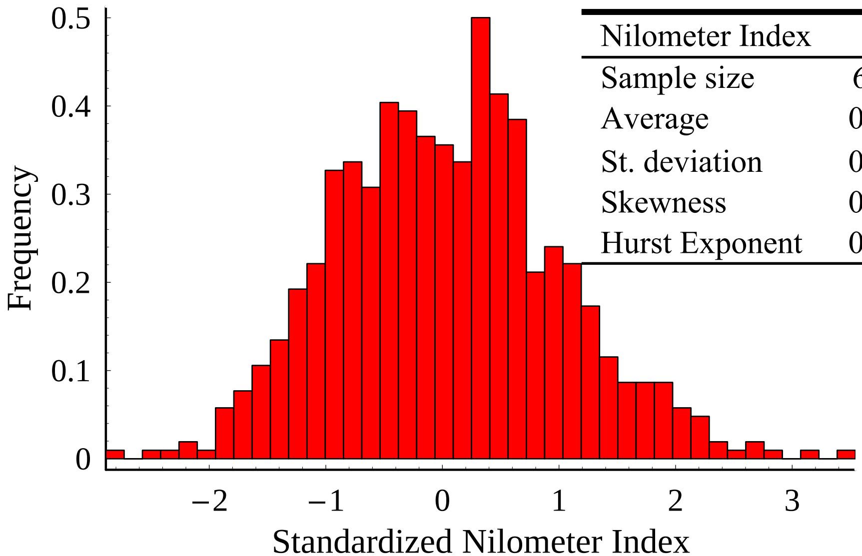
Daily Dew Points	
Sample size	3069
Average	0
St. deviation	1
Skewness	-1.10
Hurst Exponent	0.75



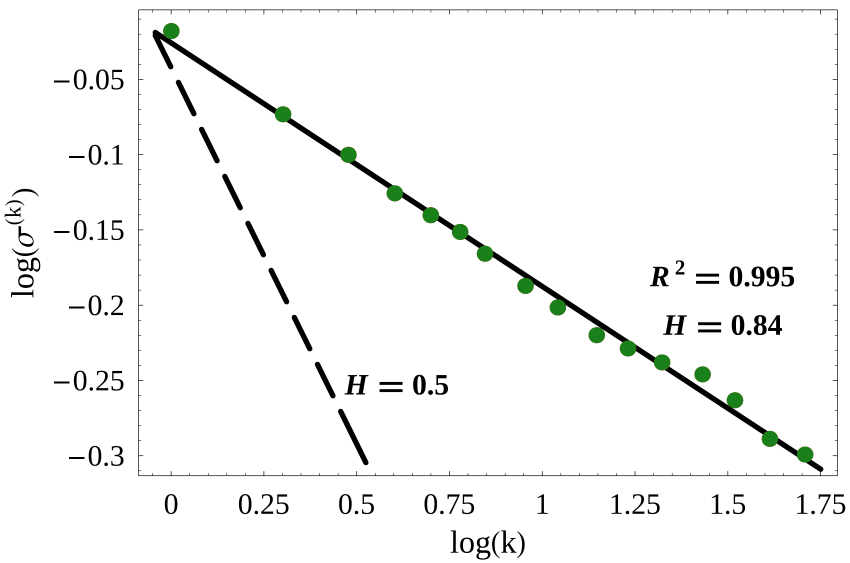
13. Simulated Data I: Nilometer Index

An auto-regressive model of order 30, AR(30), was fitted to the original dataset in order to preserve the Hurst behaviour. The skewness coefficient of the innovation was evaluated, according to the methodology analysed in this study, to $Csk_{\varepsilon} = 0.46$. A simulated series with statistics in close proximity to the observed ones is presented here.



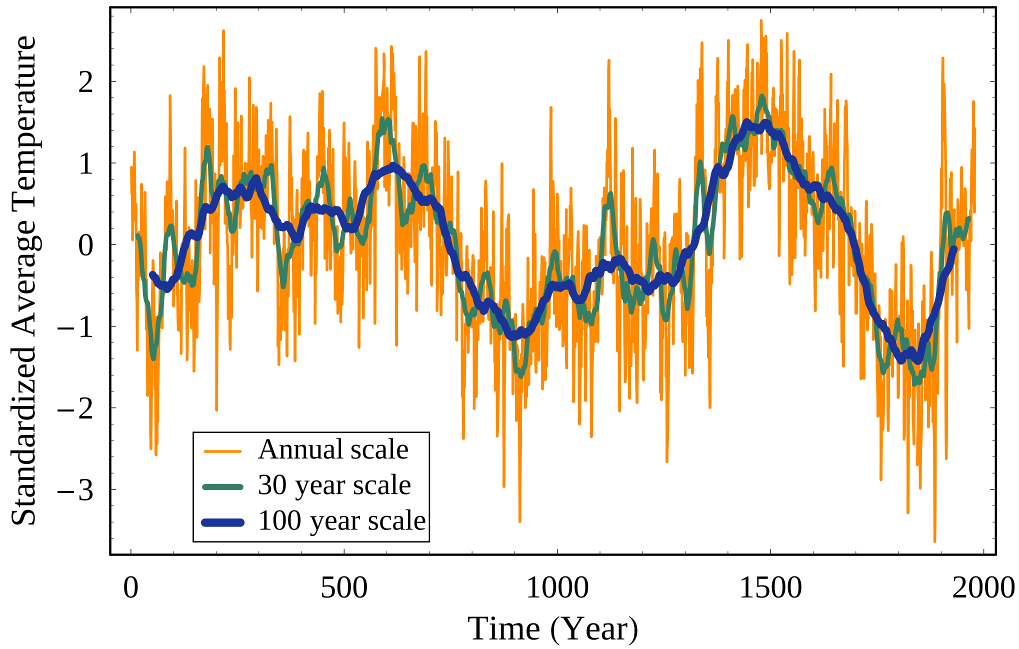


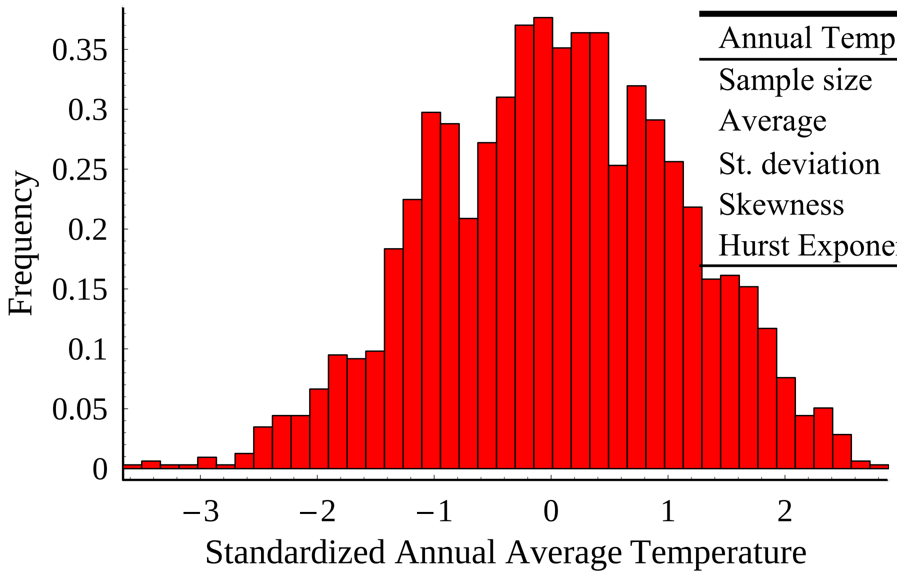
Nilometer Index	
Sample size	663
Average	0.04
St. deviation	0.96
Skewness	0.24
Hurst Exponent	0.84



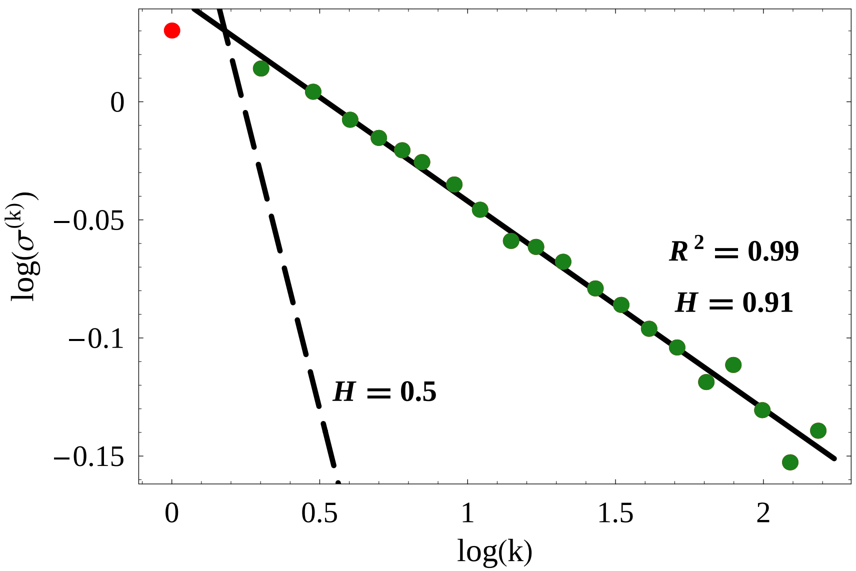
14. Simulated Data II: Global Temperatures

An auto-regressive model of order 40, AR(40), was fitted to the original dataset in order to preserve the Hurst behaviour. The skewness coefficient of the innovation was evaluated, according to the methodology analysed in this study, to $Csk_{\varepsilon} = -0.48$. A simulated series with statistics in close proximity to the observed ones is presented here.



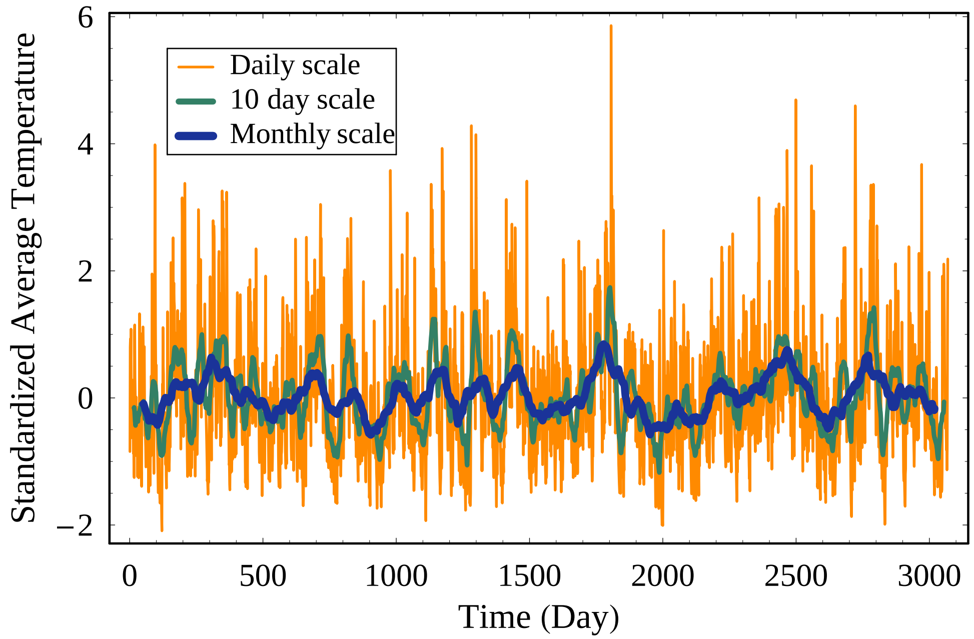


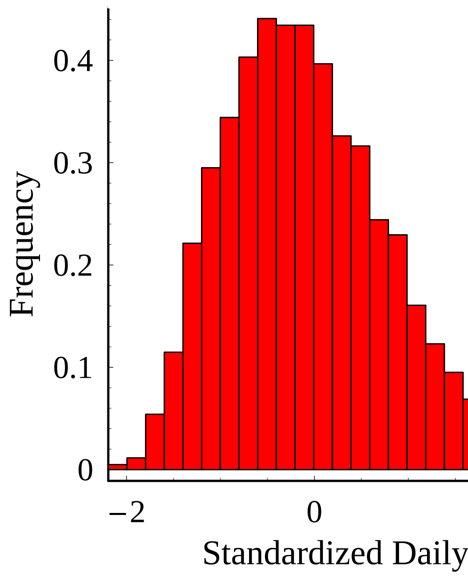
Annual Temperatures	
Sample size	1979
Average	0.04
St. deviation	1.07
Skewness	-0.12
Hurst Exponent	0.91



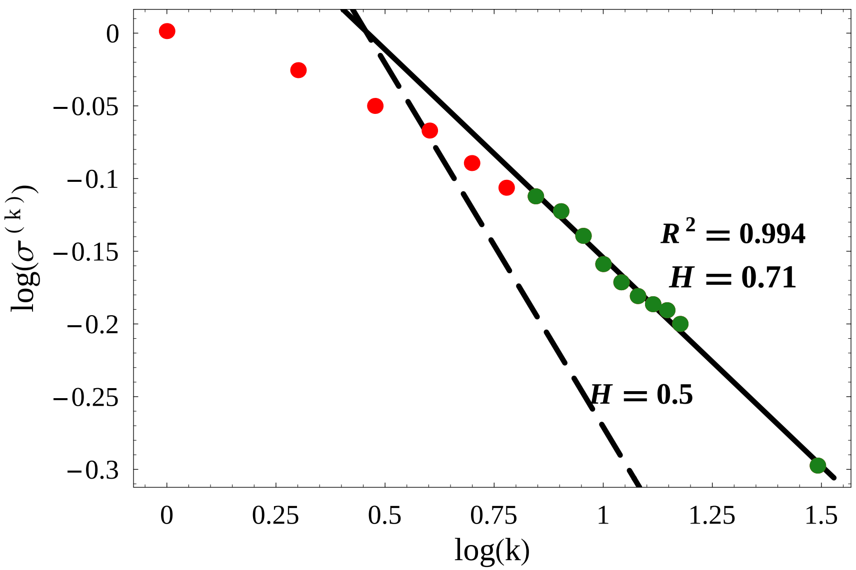
15. Simulated Data III: Daily Average Temperatures

An auto-regressive model of order 20, AR(20), was fitted to the original dataset in order to preserve the Hurst behaviour. The skewness coefficient of the innovation was evaluated, according to the methodology analysed in this study, to $Csk_{\varepsilon} = 1.94$. A simulated series with statistics in close proximity to the observed ones is presented here.



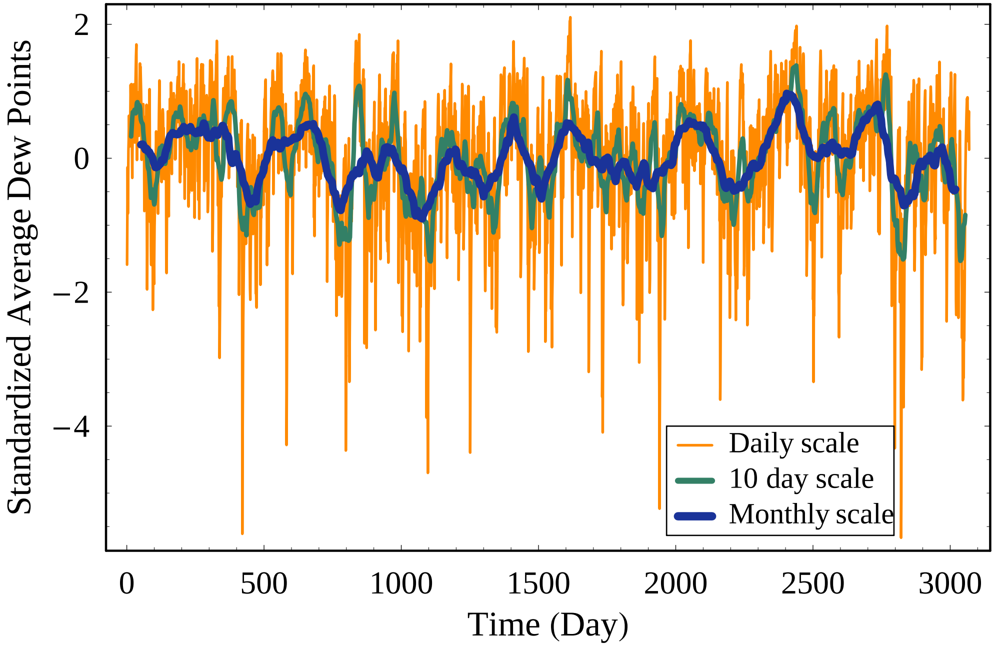


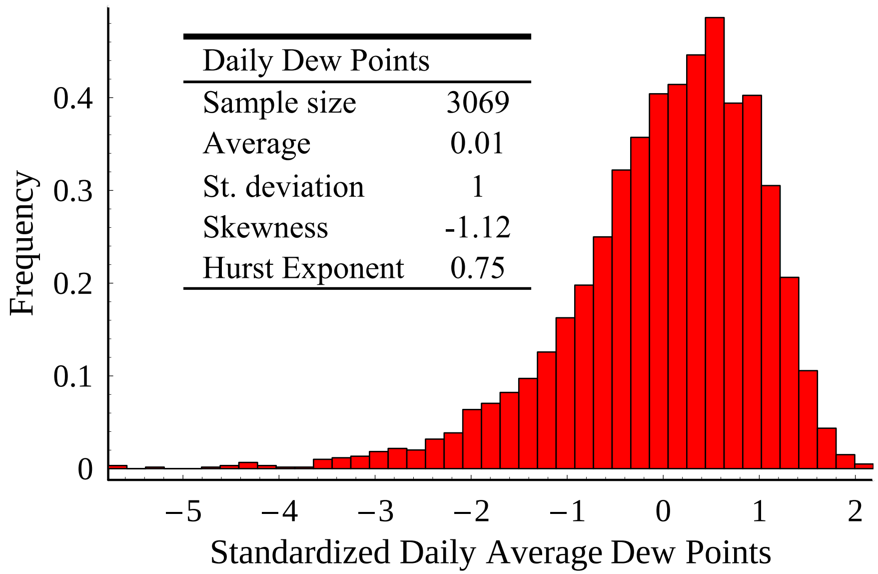
Daily Temperatures	
Sample size	3069
Average	0.02
St. deviation	1
Skewness	0.94
Hurst Exponent	0.71



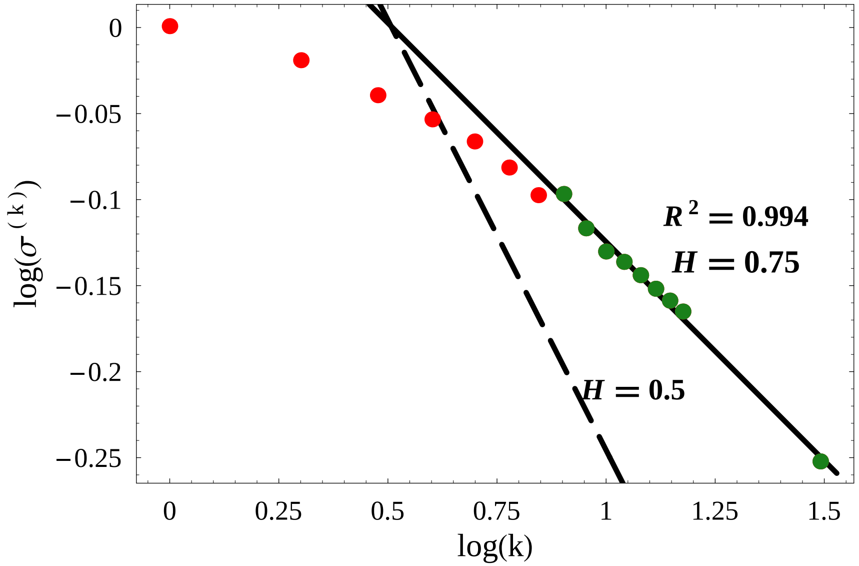
16. Simulated Data IV: Daily Average Dew Points

An auto-regressive model of order 20, AR(20), was fitted to the original dataset in order to preserve the Hurst behaviour. The skewness coefficient of the innovation was evaluated, according to the methodology analysed in this study, to Csk_{ε} = -2.65. A simulated series with statistics in close proximity to the observed ones is presented here.





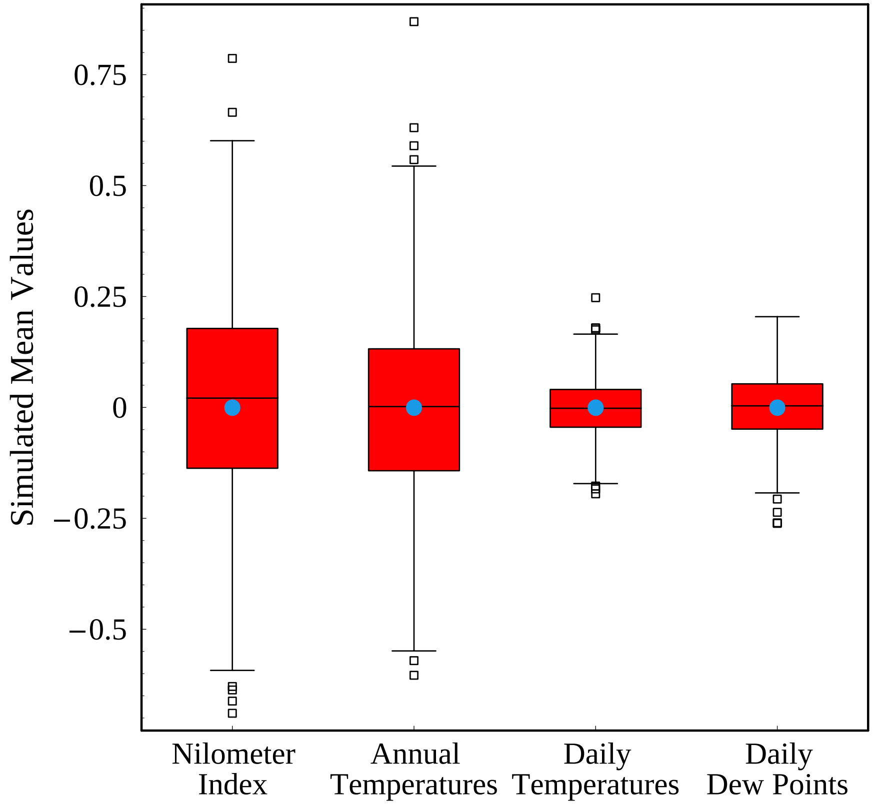
Daily Dew Points	
Sample size	3069
Average	0.01
St. deviation	1
Skewness	-1.12
Hurst Exponent	0.75

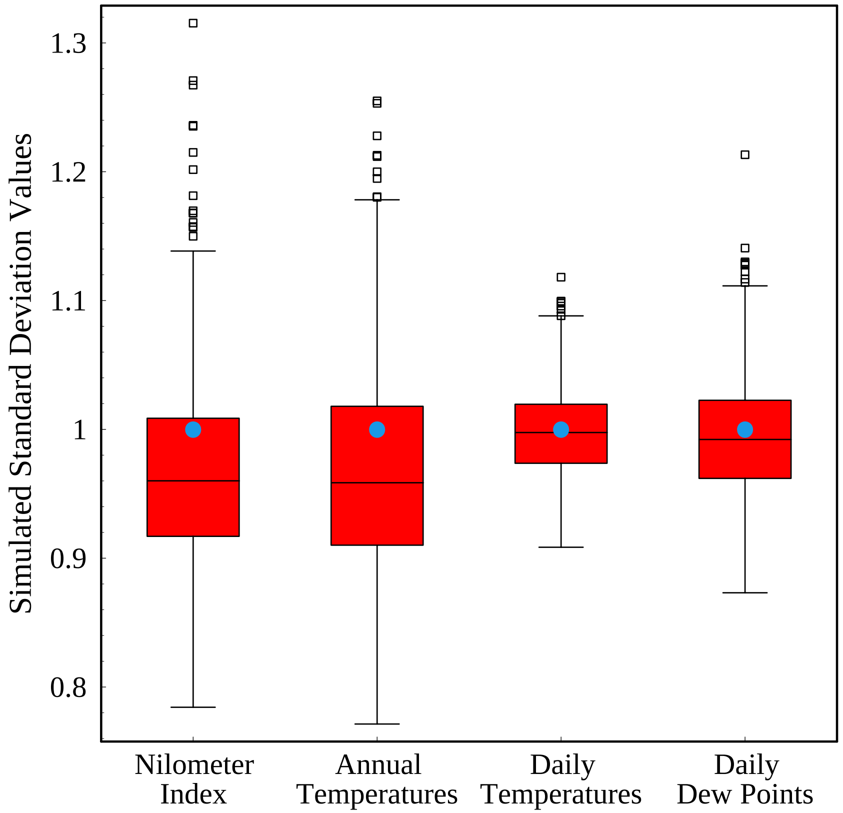


17. Simulated Mean and Standard Deviation

- The box plots below depict the estimated mean values of the 1000 simulated series. Blue dots represent the observed means of each dataset (0 as the series were standardised). The effect of the Hurst behaviour is clearly manifested by larger variability of the estimator in the series with the larger Hurst exponent values.

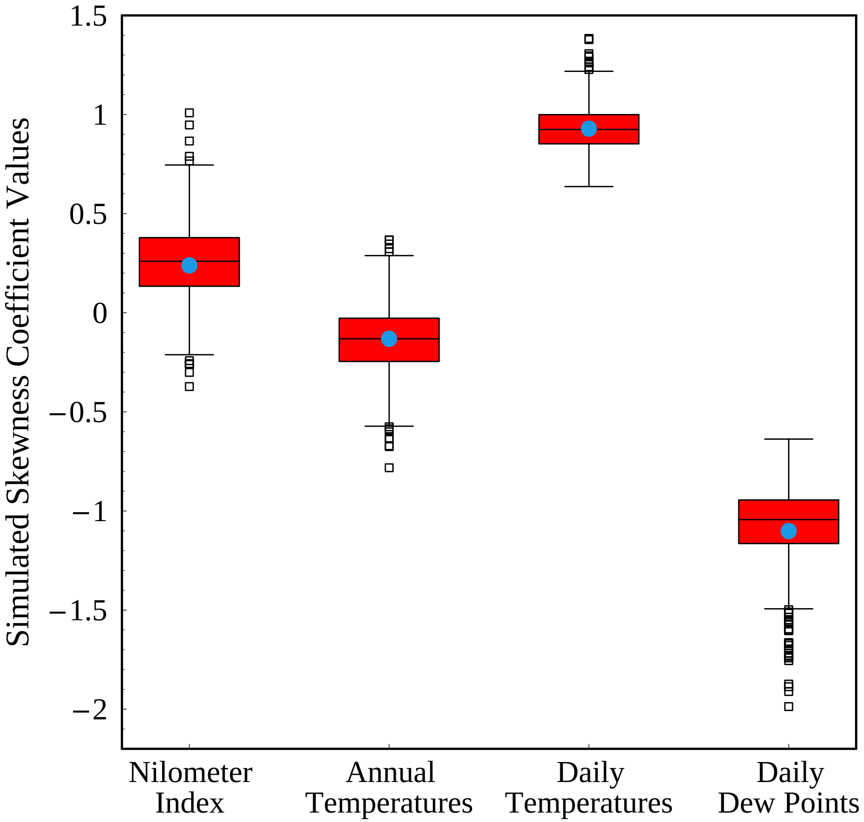
- The box plots below depict the estimated standard deviation values of the 1000 simulated series. Blue dots represent the standardized values. The variability of the classic standard deviation estimator, is larger in the series with larger values of Hurst exponent, and it is also negative biased.



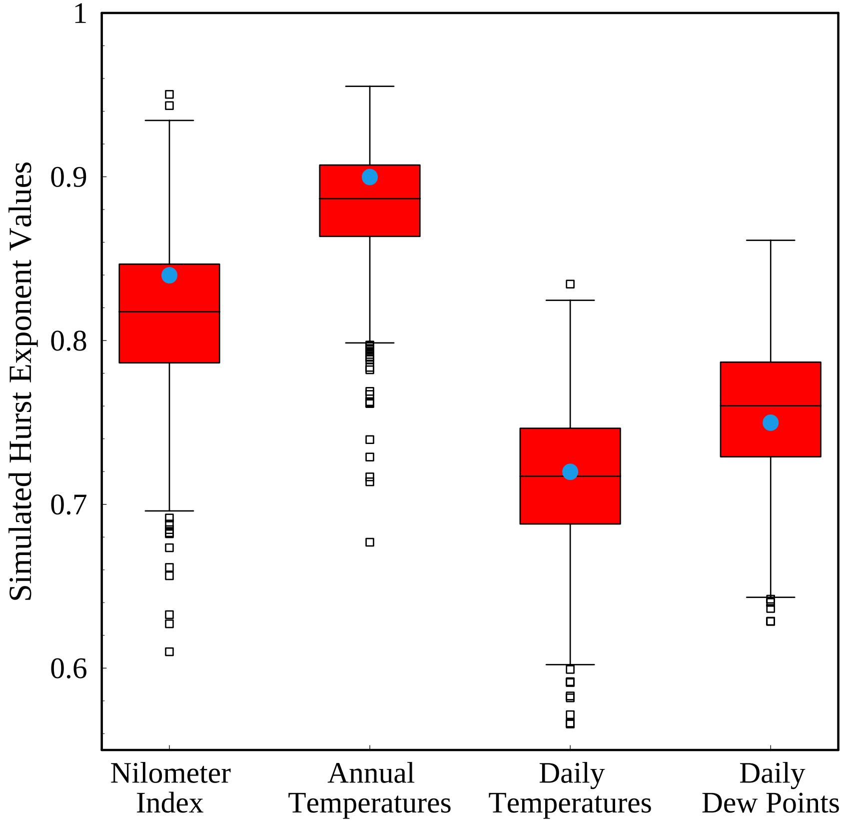


18. Simulated Skewness and Hurst Exponent

- The box plots below depict the estimated skewness coefficient values of the 1000 simulated series. Blue dots represent the observed values of each dataset. Given that the estimation of the *C*sk is highly uncertain, as its value is sensitive to outliers, it is encouraging that the model reproduces *C*sk values in proximity with the observed ones and with low variability.



- The box plots below depict the estimated Hurst exponent values of the 1000 simulated series. Blue dots represent the observed values. Again the model, as the plot reveals, manages to reproduce sufficiently the Hurst behaviour, with the mean Hurst exponent value of the simulated series in agreement with the observed ones.



19. Conclusions

- While the autoregressive models are considered to be short range persistence models, it is concluded in this study that a higher order AR model preserves adequately the Hurst behaviour, for Hurst exponent values as high as 0.9. It seems that the model can preserve even more intense long-term persistence but this needs to be further examined.
- To preserve the asymmetry, an analytical expression for the estimation of the skewness coefficient of the innovation is given. Subsequently, the innovation sequence is sampled from a flexible skewed distribution, the so-called Generalized Lambda Distribution. The model manages to preserve sufficiently the skewness as the mean skewness coefficient of the simulated series is in proximity with the observed ones.
- As the simulated series are in accordance with the observed ones, the model can be used for any practical modeling purposes.
- Overall, the proposed methodology is simple and robust.

20. References

- Beran, J., Statistics for Long Memory processes, Volume 61 of Monographs on Statistics and Applied Probability, New York: Chapman and Hall, 1994.
- Bras, R. L., and I. Rodriguez-Iturbe, Random Functions and Hydrology, 559 pp., Addison-Wesley-Longman, Reading, Mass., 1985.
- Ditlevsen, O. D., Extremes and first passage times, Doctoral dissertation, Tech. Univ. of Denmark, Lyngby, Denmark, 1971.
- Hosking, J. R. M.,Fractional Differencing, Biometrika, vol. 68, pp. 165-176, 1981.
- Hurst, H., Long-term storage capacity of reservoirs, Transactions of the American Society of Civil Engineers 116, 770−808, 1951.
- Karian, Z., Dudewicz, E., Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods, Boca Raton: CRC Press, 2000.
- Koutsoyiannis, D., A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series, *Water Resources Research*, 36(6), 1519-1533, 2000.
- Koutsoyiannis, D., The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrological Sciences Journal*, 47(4), 573-595, 2002.
- Mandelbrot, B. B., Une class de processus stochastiques homothetiques a soi: Application a la loi climatologique de H. E. Hurst, C. R. Hebd. Seances Acad. Sci., 260, 3284–3277, 1965.
- Mandelbrot, B. B., A fast fractional Gaussian noise generator, WaterResour. Res., 7(3), 543–553, 1971.
- Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises, 1, Averages and variances, Water Resour. Res., 5(1), 228–241, 1969a.
- Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises, 2, Rescaled ranges and spectra, Water Resour. Res., 5(1), 242–259, 1969b.
- Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional
- Gaussian noises, 3, Mathematical appendix, Water Resour. Res., 5(1), 260–267, 1969c.
- Matalas, N. C., and J. R. Wallis, Generation of synthetic flow sequences, in Systems Approach to Water Management, edited by A. K. Biswas, McGraw-Hill, New York, 1976.
- Moberg, A., D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karlen, Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433(7026), 613– 617, 2005.
- Ramberg, J., Schmeiser, B., An approximate method for generating asymmetric random variables, Communications of the ACM 17(2):78–82, 1974.
- Ramberg, J. S., Dudewich, E. J., Tadikamalla, P. R., Mykytka, E. F., A probability distribution and its uses in fitting data. Communications in Statistics—Theory and Methods 21(2):201–214, 1979.
- Tukey, J. W., The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts, Technical Report 36, Statistical Techniques Research Group, Princeton University,1960.