Supplementary material for the paper

Climate change, the Hurst phenomenon, and hydrological statistics

DEMETRIS KOUTSOYIANNIS

Department of Water Resources, Faculty of Civil Engineering, National Technical University, Athens

Heroon Polytechneiou 5, GR-157 80 Zographou, Greece (dk@itia.ntua.gr)

ESTIMATION OF VARIANCE AND STANDARD DEVIATION FOR KNOWN HURST COEFFICIENT

To demonstrate the consequences of using the inappropriate classic estimators of variance and standard deviation, a Monte Carlo experiment has been performed. A long series of SSS with $H = 0.8$, $\mu = 2$ and $\sigma = 0.5$ was generated. From this series, an ensemble of 100 samples each with length $n = 100$ or 50 was constructed and the sample standard deviations using both estimators were obtained. The same was done for aggregation levels $k = 1$ to 10 when $n = 100$ and $k = 1$ to 5 when $n = 50$ (so that in any aggregation level the number of items n / k is at least 10).

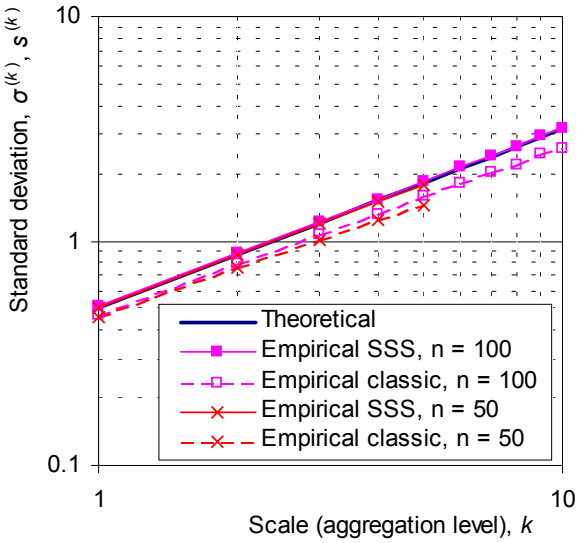

Figure A.1 Comparison of theoretical and empirical standard deviation of the aggregated processes $Z_i^{(k)}$ versus timescale k (logarithmic plots) for a Monte Carlo experiment with theoretical $H = 0.8$ and $\sigma = 0.5$.

The results are shown graphically in Figure A.1 in a logarithmic plot of standard deviation versus scale. The true standard deviation for each aggregation level k is obtained from (4). Its empirical values are obtained as the averages of the 100 samples. It is observed that, at the basic scale ($k = 1$), the classic estimators underestimate the true standard deviation by about

6% and 9% for n = 100 and 50, respectively (which is not a serious underestimation). The percentage of underestimation increases to about 20% at the largest scale used, i.e., $k = n / 10$. The SSS estimates agree perfectly with the theoretical ones (the two curves are practically indistinguishable). In addition, the classic statistics underestimate the variance of the standard deviation (not shown in Figure A.1) by 63% and 43% for n = 100 and 50, respectively.

SIMULTANEOUS ESTIMATION OF VARIANCE AND HURST COEFFICIENT

The following algorithm can be set up to determine the standard deviation σ and the Hurst exponent H that minimize the error $e^2(\sigma, H)$ in (15). Taking the derivatives of e^2 with respect to ln σ and H end equating to zero one obtains

$$\frac{1}{2}\frac{\partial e^2(\sigma, H)}{\partial \ln \sigma} = a_{11} \ln \sigma + a_{12} H - b_1(H) = 0 \tag{A.1}$$

$$\frac{1}{2}\frac{\partial e^2(\sigma, H)}{\partial H} = a_{21}(H) \ln \sigma + a_{22}(H)H - b_2(H) = 0 \tag{A.2}$$

where

$$a_{11} := \sum_{k=1}^{k'} \frac{1}{k^p},\ a_{12} := \sum_{k=1}^{k'} \frac{\ln k}{k^p},\ b_1(H) := \sum_{k=1}^{k'} \frac{\ln s^{(k)}}{k^p} - \sum_{k=1}^{k'} \frac{\ln c_k(H)}{k^p} \tag{A.3}$$

$$a_{21}(H) := \sum_{k=1}^{k'} \frac{d_k(H)}{k^p},\ a_{22}(H) := \sum_{k=1}^{k'} \frac{d_k(H) \ln k}{k^p},\ b_2(H) := \sum_{k=1}^{k'} \frac{d_k(H) \ln s^{(k)}}{k^p} - \sum_{k=1}^{k'} \frac{d_k(H) \ln c_k(H)}{k^p} \tag{A.4}$$

$$d_k(H) := \ln k + \frac{\partial \ln c_k(H)}{\partial H} = \ln k + \frac{\ln(n/k)}{1 - (n/k)^{2-2H}} \tag{A.5}$$

Eliminating ln σ it is obtained that

$$H = \frac{a_{11}\, b_2(H) - a_{21}(H)\, b_1(H)}{a_{11}\, a_{22}(H) - a_{21}(H)\, a_{12}} \tag{A.6}$$

In this equation H appears in both sides. However, it can be easily solved in an iterative manner. Assuming an initial value H = 0.5 and substituting it in the right-hand side one calculates (on the left-hand side) an improved estimate and continues this way until convergence. Having found H, σ is obtained directly from (A.1). Alternatively, it can be estimated from (11) using the standard deviation of the finest time scale only.

It can be observed that when H tends to 1, $c_k(H)$ tends to zero and most a and b terms tend to infinity, whereas for $H > 1$, $c_k(H)$ is not defined. Therefore, the method will never result in values of $H > 1$. Also, from (A.1) it becomes clear that when H tends to 1, σ tends to infinity, which is expected (from (10)). Although this behaviour is theoretically consistent, it may result in unreasonably high variables when the estimated H is higher than, say 0.98. Such

high values are not met in hydrological time series, but to ensure avoiding them, a penalty factor $H^{q+1}/(q+1)$ could be added to e^2 in (15) for a high q, say 50. This will result in an additional term equal to $-H^q$ in $b_2(H)$ in (A.4).

Some information on the behaviour of the proposed algorithm, additional to that of Figure 4 is provided in Figure A.2, where the estimated standard deviation is plotted against the estimated Hurst exponent. The figure indicates that for $H < 0.85$ the two statistics are practically uncorrelated but for higher H they become positively correlated.

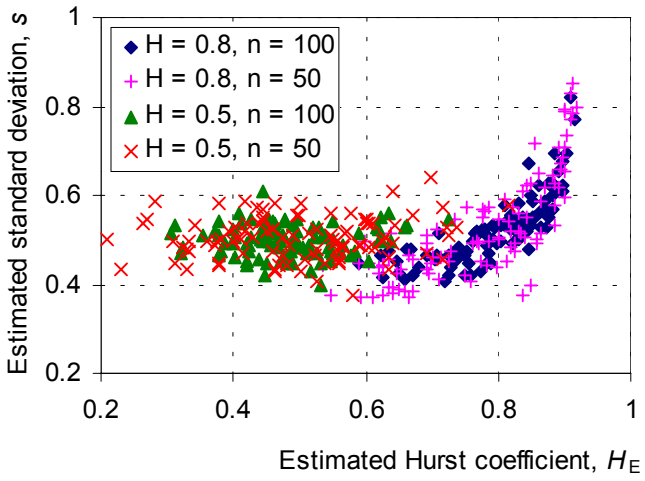

Figure A.2 Estimated Hurst coefficients versus estimated standard deviations from the ensembles of synthetic series of the Monte Carlo experiment of Figure 4 and for the proposed estimation method.

ESTIMATION OF CROSS-COVARIANCES AND CROSS-CORRELATIONS

Assuming that two processes X_i and Y_i are both SSS with common H and mutually correlated, the typical covariance estimator

$$S_{XY} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{A.7}$$

is a biased estimator. To show this, (A.7) is rewritten as

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \mu_X) - (\overline{X} - \mu_X)]\,[(Y_i - \mu_Y) - (\overline{Y} - \mu_Y)] \tag{A.8}$$

and, after algebraic manipulations, also considering (9), it is obtained that

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)\,(Y_i - \mu_Y) - \frac{n}{n-1}(\overline{X} - \mu_X)\,(\overline{Y} - \mu_Y) \tag{A.10}$$

Taking expected values in (A.10) and also assuming, by analogy to (11), that the aggregated covariance is

$$\mathrm{Cov}[X_1 + \ldots + X_n,\, Y_1 + \ldots + Y_n] = n^{2H}\,\mathrm{Cov}[X_i,\, Y_i] \tag{A.12}$$

one obtains

$$E[S_{XY}] = \frac{n - n^{2H-1}}{n-1}\,\mathrm{Cov}[X_i,\, Y_i] \tag{A.13}$$

which proves that (A.7) is unbiased only when $H = 0.5$. Consequently, the SSS unbiased covariance estimator for any known H is

$$\widetilde{S}_{XY} := \frac{n-1}{n - n^{2H-1}}\,S_{XY} = \frac{1}{n - n^{2H-1}} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \tag{A.14}$$

Here, it is observed that the correlation coefficient is

$$R_{XY} := \frac{S_{XY}}{S_X\, S_Y} = \frac{\widetilde{S}_{XY}}{\widetilde{S}_X\, \widetilde{S}_Y} \tag{A.15}$$

Thus, the classic estimator of the cross-correlation coefficient remains valid also for SSS.

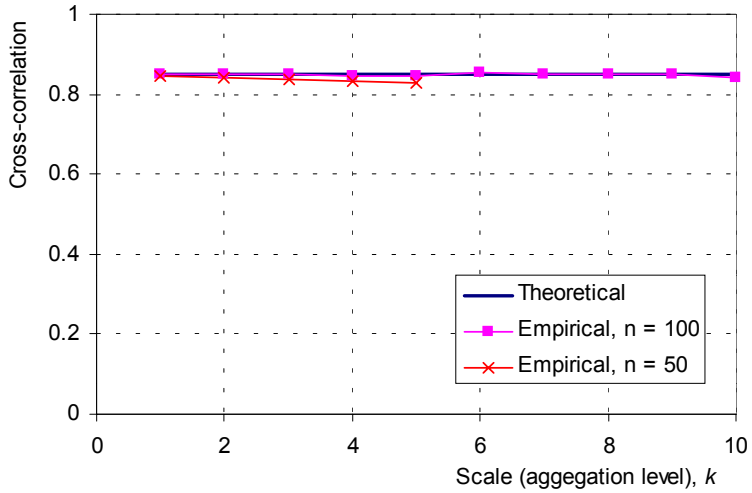

Figure A.3 Comparison of theoretical and empirical cross-correlation coefficients of a bivariate aggregated process $\mathbf{Z}_i^{(k)}$ versus timescale k for a Monte Carlo experiment with theoretical Hurst coefficient 0.8 for both variates, standard deviations 0.5 and 1.2 for the first and second variate, respectively, and theoretical cross-correlation coefficient 0.85.

This is demonstrated in Figure A.3. A Monte Carlo experiment was performed here by generating bivariate synthetic samples with common $H = 0.8$ and length $n = 100$ or 50. The other characteristic parameters were $\mu = 2$ and 3, and $\sigma = 0.5$ and 1.2 for the first and second variable, respectively. The theoretical cross-correlation coefficient was 0.85. To generate the bivariate synthetic samples the multivariate method by *Koutsoyiannis* (2000) was followed. From the ensembles of 100 series, the empirical cross-correlations for several timescales were estimated, which, as shown in Figure A.3, agree well with the theoretical expectation.

ESTIMATION OF AUTO-COVARIANCES AND AUTO-CORRELATIONS

The typical estimator of the lag l autocovariance (Equation (20)) can be written as

$$G_l = \frac{1}{n}\sum_{i=1}^{n-l}[(X_i-\mu)-(\overline{X}-\mu)]\,[(X_{i+l}-\mu)-(\overline{X}-\mu)] \tag{A.16}$$

After algebraic manipulations, also considering (17), one obtains

$$G_l = \frac{1}{n}\sum_{i=1}^{n-l}(X_i-\mu)\,(X_{i+l}-\mu) - \frac{1}{n}(\overline{X}-\mu)\sum_{i=1}^{n-l}[(X_i-\mu)+(X_{i+l}-\mu)] + (\overline{X}-\mu)^2 \tag{A.18}$$

Assuming that l is small in comparison with n so that $n - l$ and n can be interchanged, and also the second sum of (A.18) can be extended over all i, one obtains

$$G_l \approx \frac{1}{n}\sum_{i=1}^{n-l}(X_i-\mu)\,(X_{i+l}-\mu) - (\overline{X}-\mu)^2 \tag{A.19}$$

For the SSS case, taking expected values (and again ignoring the difference of $n - l$ and n), it is found that

$$E[G_l] \approx \gamma_l - \frac{\sigma^2}{n^{2-2H}} \tag{A.20}$$

This means that an approximately unbiased estimator of γ_l is given by (21) and an approximately unbiased of the autocorrelation coefficient ρ_l is given by (22).

CASE STUDIES

Figure A.4 depicts the autocorrelation coefficients versus scale (panel (a)) and versus lag (panel (b)) Similarly to Figure 8, this figure verifies the presence of long-term persistence, the appropriateness of the proposed SSS estimator of autocorrelation, and the large departure of the classic estimations from SSS estimators.

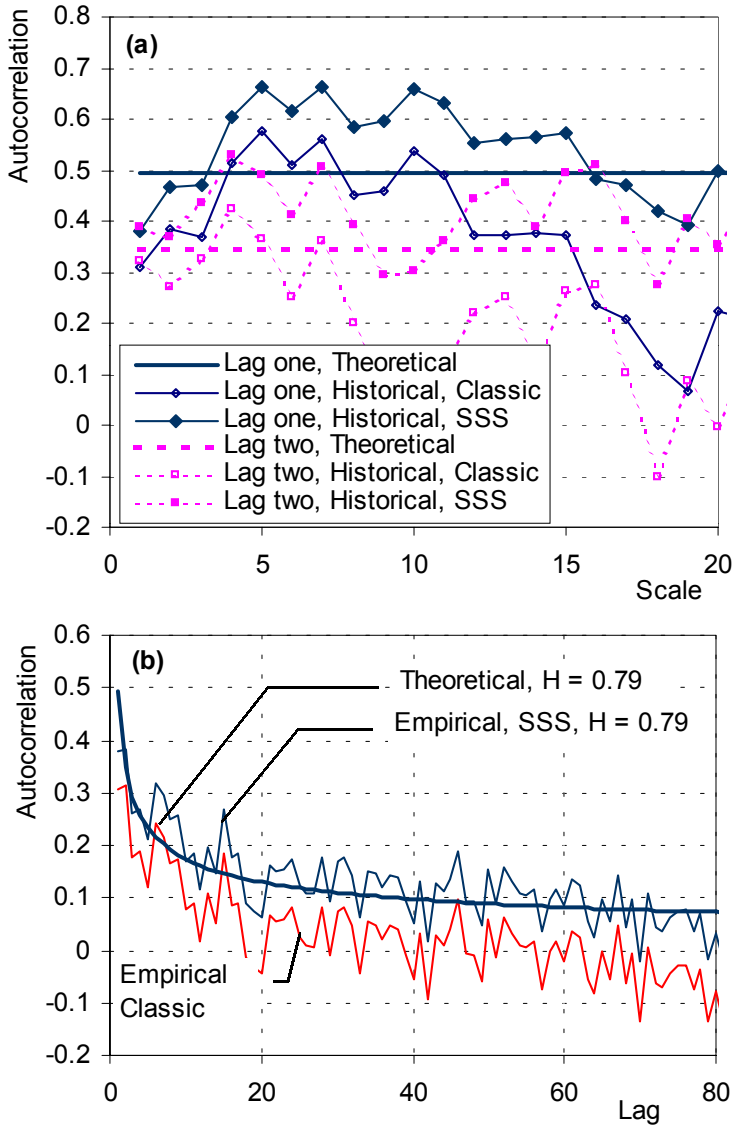

Figure A.4 Autocorrelation coefficients of the time series of mean annual temperature at Paris/Le Bourget: (a) lag 1 and lag 2 autocorrelations of the aggregated process versus timescale, k; (b) autocorrelation versus lag for the basic (annual) timescale, $k = 1$.