Supplementary Material: A quick gap-filling of missing hydrometeorological data

Christoforos Pappas[1,2], Simon Michael Papalexiou[2], Demetris Koutsoyiannis[2]

[1]Institute of Environmental Engineering, ETH Zurich, 8093 Zurich, Switzerland (pappas@ifu.baug.ethz.ch)

[2]Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece

1. Proof of Equation (5)

The error of an estimated missing variable at time t is defined as the difference between the real variable $\underline{x}_t$ and the estimate $\hat{\underline{x}}_t$. In the Optimal Local Average (OLA) methodology, a missing variable is estimated as $\hat{\underline{x}}_t = \left(\sum_{i=1}^{n} \underline{x}_{t-i} + \sum_{i=1}^{n} \underline{x}_{t+i}\right) / 2n$ where $2n$ is the number of time-adjacent values used for the infilling (i.e., n neighboring values before, and n after the missing observation). The squared error of the estimate is then given by:

$$\underline{e}^2 := \left(\underline{x}_t - \hat{\underline{x}}_t\right)^2 = \left(\underline{x}_t - \frac{\sum_{i=1}^{n} \underline{x}_{t-i} + \sum_{i=1}^{n} \underline{x}_{t+i}}{2n}\right)^2 = \underline{x}_t^{\,2} - 2\underline{x}_t \frac{\sum_{i=1}^{n} \underline{x}_{t-i} + \sum_{i=1}^{n} \underline{x}_{t+i}}{2n} + \left(\frac{\sum_{i=1}^{n} \underline{x}_{t-i} + \sum_{i=1}^{n} \underline{x}_{t+i}}{2n}\right)^2$$

$$= \underline{x}_t^{\,2} - \frac{1}{n}\underline{x}_t \sum_{i=1}^{n} \underline{x}_{t-i} - \frac{1}{n}\underline{x}_t \sum_{i=1}^{n} \underline{x}_{t+i} + \frac{\left(\sum_{i=1}^{n} \underline{x}_{t-i}\right)^2 + \left(\sum_{i=1}^{n} \underline{x}_{t+i}\right)^2 + 2\sum_{i=1}^{n} \underline{x}_{t-i} \sum_{i=1}^{n} \underline{x}_{t+i}}{4n^2}$$

$$= \underline{x}_t^{\,2} - \frac{1}{n}\underline{x}_t \sum_{i=1}^{n} \underline{x}_{t-i} - \frac{1}{n}\underline{x}_t \sum_{i=1}^{n} \underline{x}_{t+i} + \frac{1}{4n^2}\left(\sum_{i=1}^{n} \underline{x}_{t-i}\right)^2 + \frac{1}{4n^2}\left(\sum_{i=1}^{n} \underline{x}_{t+i}\right)^2 + \frac{1}{2n^2}\sum_{i=1}^{n} \underline{x}_{t-i} \sum_{i=1}^{n} \underline{x}_{t+i}$$

The expected value of the squared error is the MSE of the estimation and it can be expressed as:

$$\mathrm{MSE} \coloneqq \mathrm{E}\left[\underline{e}^2\right] = \mathrm{E}\left[\underline{x}_t^{\,2} - \frac{1}{n}\underline{x}_t\sum_{i=1}^{n}\underline{x}_{t-i} - \frac{1}{n}\underline{x}_t\sum_{i=1}^{n}\underline{x}_{t+i} + \frac{1}{4n^2}\left(\sum_{i=1}^{n}\underline{x}_{t-i}\right)^2 + \frac{1}{4n^2}\left(\sum_{i=1}^{n}\underline{x}_{t+i}\right)^2 + \frac{1}{2n^2}\sum_{i=1}^{n}\underline{x}_{t-i}\sum_{i=1}^{n}\underline{x}_{t+i}\right]$$

$$= \mathrm{E}\left[\underline{x}_t^{\,2}\right] - \frac{1}{n}\mathrm{E}\left[\underline{x}_t\sum_{i=1}^{n}\underline{x}_{t-i}\right] - \frac{1}{n}\mathrm{E}\left[\underline{x}_t\sum_{i=1}^{n}\underline{x}_{t+i}\right] + \frac{1}{4n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n}\underline{x}_{t-i}\right)^2\right] + \frac{1}{4n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n}\underline{x}_{t+i}\right)^2\right] + \frac{1}{2n^2}\mathrm{E}\left[\sum_{i=1}^{n}\underline{x}_{t-i}\sum_{i=1}^{n}\underline{x}_{t+i}\right]$$

Assuming that the underlying process is stationary with mean μ, standard deviation σ, and correlation coefficient for lag i ρ_i, following basic rules of statistics we obtain:

$$\mathrm{E}\left[\underline{x}_t^{\,2}\right] = \sigma^2 + \mu^2$$

$$\mathrm{E}\left[\frac{1}{n}\underline{x}_t\sum_{i=1}^{n}\underline{x}_{t-i}\right] = \mathrm{E}\left[\frac{1}{n}\underline{x}_t\sum_{i=1}^{n}\underline{x}_{t+i}\right] = \frac{1}{n}\sigma^2\sum_{i=1}^{n}\rho_i + n\mu^2$$

$$\mathrm{E}\left[\frac{1}{4n^2}\left(\sum_{i=1}^{n}\underline{x}_{t-i}\right)^2\right] = \mathrm{E}\left[\frac{1}{4n^2}\left(\sum_{i=1}^{n}\underline{x}_{t+i}\right)^2\right] = \frac{1}{4n^2}\left[\sigma^2\left(n + 2\sum_{i=1}^{n-1}(n-i)\rho_i\right) + n^2\mu^2\right]$$

$$\mathrm{E}\left[\frac{1}{2n^2}\sum_{i=1}^{n}\underline{x}_{t-i}\sum_{i=1}^{n}\underline{x}_{t+i}\right] = \frac{1}{2n^2}\left[\sigma^2\left(\sum_{i=2}^{n+1}(i-1)\rho_i + \sum_{i=n+2}^{2n}(2n+1-i)\rho_i\right) + n^2\mu^2\right]$$

The MSE can then be written as:

$$\mathrm{MSE} \coloneqq \mathrm{E}\left[\underline{e}^2\right] =$$

$$= \sigma^2 + \mu^2 - \frac{2}{n}\sigma^2\sum_{i=1}^{n}\rho_i + n\mu^2 + \frac{2}{4n^2}\left[\sigma^2\left(n + 2\sum_{i=1}^{n-1}(n-i)\rho_i\right) + n^2\mu^2\right]$$

$$+ \frac{1}{2n^2}\left[\sigma^2\left(\sum_{i=2}^{n+1}(i-1)\rho_i + \sum_{i=n+2}^{2n}(2n+1-i)\rho_i\right) + n^2\mu^2\right]$$

And after some algebraic simplifications:

$$\mathrm{MSE} \coloneqq \mathrm{E}\left[\underline{e}^2\right] = \frac{1}{2}\left(\frac{\sigma}{n}\right)^2\left[(2n+1)\left(n - 2\sum_{i=1}^{n}\rho_i\right) + \sum_{i=1}^{2n}(2n+1-i)\,\rho_i\right]$$

A Monte Carlo confirmation of the relationship between MSE and lag-1 autocorrelation is illustrated in Figure S1. Figure S2 provides also an additional illustration of the Eq. (5) for the two examined autocorrelation structures.

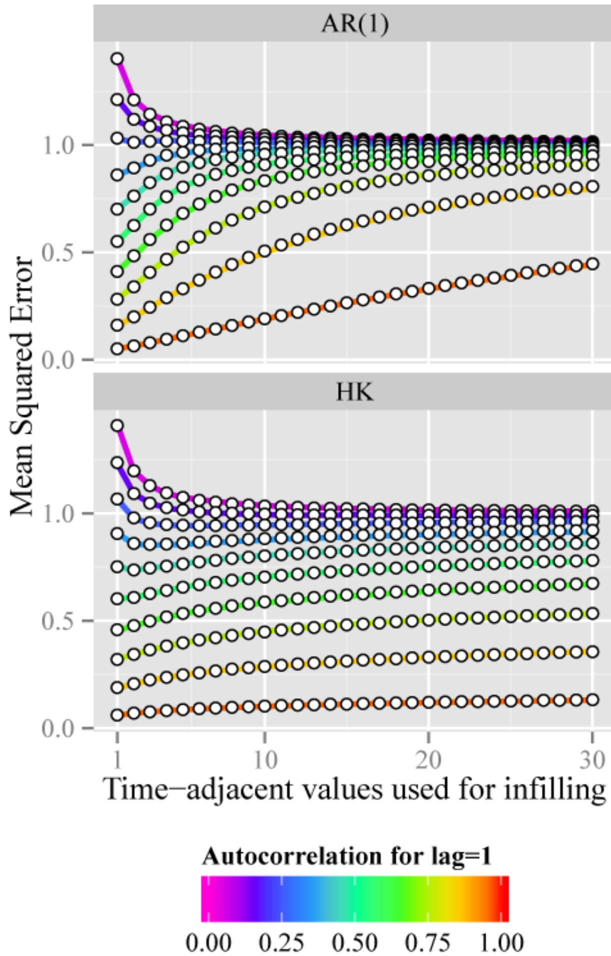

Figure S1. Monte Carlo confirmation of Eq. (5). Solid lines represent the Mean Squared Error (MSE) as estimated by Eq. (5), while the points correspond to the calculated MSE from the Monte Carlo simulations. Time series with 100000 values were generated from AR(1) and HK processes with zero mean and standard deviation equal to one and various values of lag-1 autocorrelation coefficient. The time series with HK dynamics were simulated using the function *SimulateFGN* from the `R` package FGN (Veenstra & McLeod, 2012).

2. Optimal Local Average (OLA) additional material

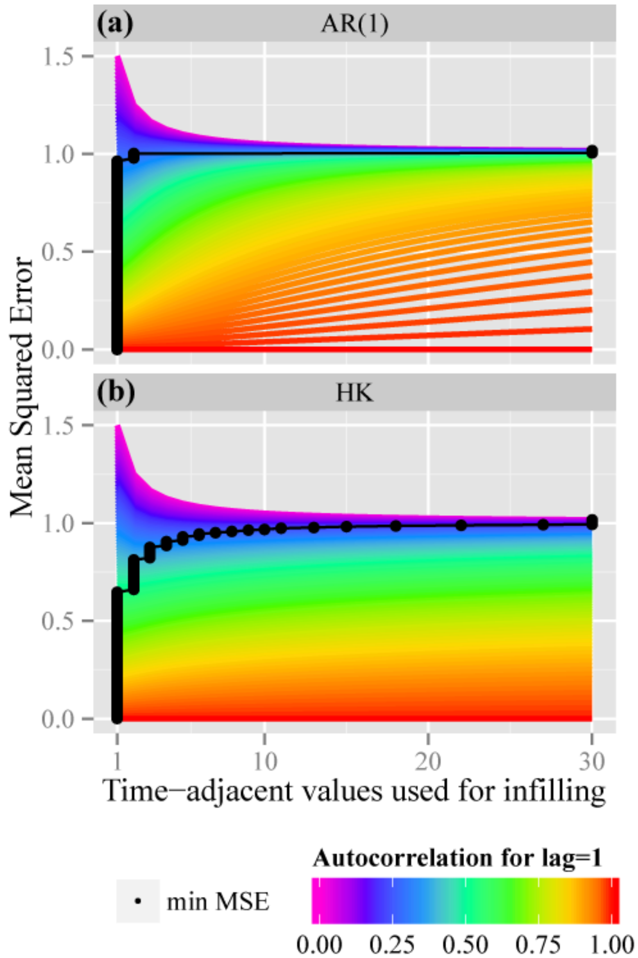

Figure S2. Surface plots illustrating the Optimal Local Average (OLA) methodology, based on Eq. (5) with $\sigma = 1$, for processes with exponential (a), and with power-law (b) autocorrelation structure. For a wide range of lag-1 autocorrelations, for both structures, the optimal infilling, i.e., minimum Mean Squared Error (MSE) occurs when a local average is used, instead for the commonly used sample (global) average (depicted above with 30 time-adjacent values). For lag-1 autocorrelation greater than 0.52, for both the examined autocorrelation structures, the strictly local average (i.e., by using one value before and one after the missing record) provides the best results (minimum MSE) while the use of sample average inflates the MSE.

3. Proof of Equation (8)

The error of an estimated missing value at time t is defined as the difference between the real value of the variable $\underline{x}_t$ and the estimated value $\hat{\underline{x}}_t$. When the Weighted Sum of local and total Average (WSA) is applied, the missing variable is estimated as $\hat{\underline{x}}_t = \lambda\left(\sum_{i=1}^{N}(\underline{x}_{t-i}+\underline{x}_{t+i})\right)/2N + (1-\lambda)\left(\sum_{i=1}^{n}\underline{x}_{t-i}+\sum_{i=1}^{n}\underline{x}_{t+i}\right)/2n$ where N is the number of available observations before (or after) the missing values, corresponding to the global average, n is the range of the local average (i.e., the number of time-adjacent values used for the infilling) and λ is a factor (weight) regulating the contribution of the global (i.e., $\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)/2N$) and the local (i.e., $\left(\sum_{i=1}^{n}\underline{x}_{t-i}+\sum_{i=1}^{n}\underline{x}_{t+i}\right)/2n$) average. Since the methodology is developed envisioning fast and direct applicability, the local average is restricted to only one neighboring value (i.e., n = 1, one value before, and one after the missing observation). Therefore, the missing value is estimated as $\underline{\hat{x}}_t=\lambda\left(\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)\right)/2N+(1-\lambda)\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)/2$. The squared error of the estimate is then given by:

$$
\begin{aligned}
\underline{e}^2 &:= \left(\underline{x}_t-\underline{\hat{x}}_t\right)^2=\left[\underline{x}_t-\left(\lambda\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}+(1-\lambda)\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)\right]^2\\
&=\left[\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)-\lambda\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)\right]^2
\end{aligned}
$$

$$
=\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2-2\lambda\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)+\lambda^2\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2
$$

For the sake of readability, we separate the following quantities:

$$
\underline{A}=\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2,\underline{B}=\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right),\underline{C}=\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2
$$

The squared error can be then summarized as $\underline{e}^2=\underline{A}-2\lambda\underline{B}+\lambda^2\underline{C}$ and the Mean Squared Error (MSE), $\mathrm{E}\left[\underline{e}^2\right]$, is given by $\mathrm{MSE}:=\mathrm{E}\left[\underline{e}^2\right]=\mathrm{E}\left[\underline{A}\right]-2\lambda\mathrm{E}\left[\underline{B}\right]+\lambda^2\mathrm{E}\left[\underline{C}\right]$. Assuming that the underlying process is stationary with mean μ, standard deviation σ, and correlation coefficient for lag i ρ_i, we have for each quantity:

$$\mathrm{E}\left[\underline{A}\right]=\mathrm{E}\left[\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2\right]$$

But from Eq. (5) we have proven that

$$\mathrm{E}\left[\left(\underline{x}_t-\frac{\sum_{i=1}^{n}\underline{x}_{t-i}+\sum_{i=1}^{n}\underline{x}_{t+i}}{2n}\right)^2\right]=\frac{1}{2}\left(\frac{\sigma}{n}\right)^2\left[(2n+1)\left(n-2\sum_{i=1}^{n}\rho_i\right)+\sum_{i=1}^{2n}(2n+1-i)\rho_i\right]$$

Therefore, for n = 1, $\mathrm{E}\left[\underline{A}\right]$ can be written as

$$\mathrm{E}\left[\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2\right]=\mathrm{E}\left[\underline{A}\right]=\frac{1}{2}\sigma^2(3-4\rho_1+\rho_2)$$

$$\underline{B}=\left(\underline{x}_t-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)=$$

$$=\frac{1}{2N}\underline{x}_t\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)-\frac{1}{2}\underline{x}_t\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)-\frac{1}{4N}\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)+\frac{1}{4}\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)^2$$

We examine each term separately:

$$\underline{x}_t\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)=\underline{x}_t\sum_{i=1}^{N}\underline{x}_{t-i}+\underline{x}_t\sum_{i=1}^{N}\underline{x}_{t+i}$$

Based on algebraic manipulations similar to the ones presented in S1 we have:

$$\mathrm{E}\left[\underline{x}_t\sum_{i=1}^{N}\underline{x}_{t-i}\right]=\mathrm{E}\left[\underline{x}_t\sum_{i=1}^{N}\underline{x}_{t+i}\right]=\sigma^2\sum_{i=1}^{N}\rho_i+N\mu^2\Rightarrow\mathrm{E}\left[\underline{x}_t\sum_{i=-N}^{N}\underline{x}_{t+i}\right]=2\sigma^2\sum_{i=1}^{N}\rho_i+2N\mu^2$$

$$\mathrm{E}\left[\underline{x}_t\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)\right]=2\sigma^2\rho_1+2\mu^2$$

$$\mathrm{E}\left[\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)\right]=2\sigma^2\left(\sum_{i=1}^{N-1}\rho_i+\sum_{i=2}^{N+1}\rho_i+1\right)+4N\mu^2$$

$$\mathrm{E}\left[\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)^2\right]=2\sigma^2\left(\rho_2+1\right)+4\mu^2$$

By combining the abovementioned quantities we obtain:

$$\mathrm{E}\left[\underline{B}\right]=\sigma^2\left[\frac{1}{N}\sum_{i=1}^{N}\rho_i-\frac{1}{2N}\left(\sum_{i=1}^{N-1}\rho_i+\sum_{i=2}^{N+1}\rho_i+1\right)-\rho_1+\frac{1}{2}\rho_2+\frac{1}{2}\right]$$

$$\underline{C}=\left(\frac{\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)}{2N}-\frac{\underline{x}_{t-1}+\underline{x}_{t+1}}{2}\right)^2=\frac{\left(\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)\right)^2}{4N^2}+\frac{\underline{x}_{t-1}^2+\underline{x}_{t+1}^2+2\underline{x}_{t-1}\underline{x}_{t+1}}{4}-\frac{1}{2N}\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)$$

where

$$\left(\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)\right)^2=\left(\sum_{i=1}^{N}\underline{x}_{t-i}+\sum_{i=1}^{N}\underline{x}_{t+i}\right)^2=\left(\sum_{i=1}^{N}\underline{x}_{t-i}\right)^2+\left(\sum_{i=1}^{N}\underline{x}_{t+i}\right)^2+2\sum_{i=1}^{N}\underline{x}_{t-i}\sum_{i=1}^{N}\underline{x}_{t+i}$$

$$\begin{aligned}\mathrm{E}\left[\left(\sum_{i=1}^{N}\left(\underline{x}_{t-i}+\underline{x}_{t+i}\right)\right)^2\right]&=2\left[\sigma^2\left(N+2\sum_{i=1}^{N-1}\left(N-i\right)\rho_i\right)+N^2\mu^2\right]+2\left[\sigma^2\left(\sum_{i=2}^{N+1}(i-1)\rho_i+\sum_{i=N+2}^{2N}\left(2N+1-i\right)\rho_i\right)+N^2\mu^2\right]\\&=2\sigma^2\left[N+2\sum_{i=1}^{N-1}\left(N-i\right)\rho_i+\sum_{i=2}^{N+1}(i-1)\rho_i+\sum_{i=N+2}^{2N}\left(2N+1-i\right)\rho_i\right]+4N^2\mu^2\end{aligned}$$

$$\mathrm{E}\left[\frac{\underline{x}_{t-1}^2+\underline{x}_{t+1}^2+2\underline{x}_{t-1}\underline{x}_{t+1}}{4}\right]=\frac{\sigma^2}{2}(\rho_2+1)+\mu^2$$

$$\mathrm{E}\left[\sum_{i=-N}^{N}\underline{x}_{t+i}\left(\underline{x}_{t-1}+\underline{x}_{t+1}\right)\right]=2\sigma^2\left(\sum_{i=1}^{N-1}\rho_i+\sum_{i=2}^{N+1}\rho_i+1\right)+4N\mu^2$$

$\mathrm{E}\left[\underline{C}\right]$ can be then written as

$$\begin{aligned}\mathrm{E}\left[\underline{C}\right]&=\frac{1}{2}\left(\frac{\sigma}{N}\right)^2\left(2\sum_{i=1}^{N-1}\left(N-i\right)\rho_i+\sum_{i=2}^{N+1}\left(i-1\right)\rho_i+\sum_{i=N+2}^{2N}\left(2N+1-i\right)\rho_i+N\right)+\frac{\sigma^2}{2}\left(\rho_2+1\right)-\frac{\sigma^2}{N}\left(\sum_{i=1}^{N-1}\rho_i+\sum_{i=2}^{N+1}\rho_i+1\right)\\&=\sigma^2\left[\frac{1}{2N^2}\left(2\sum_{i=1}^{N-1}\left(N-i\right)\rho_i+\sum_{i=2}^{N+1}\left(i-1\right)\rho_i+\sum_{i=N+2}^{2N}\left(2N+1-i\right)\rho_i+N\right)+\frac{\rho_2}{2}+\frac{1}{2}-\frac{1}{N}\left(\sum_{i=1}^{N-1}\rho_i+\sum_{i=2}^{N+1}\rho_i+1\right)\right]\end{aligned}$$

Summarizing the previous quantities

$$\mathrm{MSE}:=\mathrm{E}\left[\underline{e}^2\right]=\mathrm{E}\left[\underline{A}\right]-2\lambda\mathrm{E}\left[\underline{B}\right]+\lambda^2\mathrm{E}\left[\underline{C}\right]=$$

$$= \frac{1}{2}\sigma^2(3-4\rho_1+\rho_2) - 2\lambda\sigma^2\left[\frac{1}{N}\sum_{i=1}^{N}\rho_i - \frac{1}{2N}\left(\sum_{i=1}^{N-1}\rho_i - \sum_{i=2}^{N+1}\rho_i + 1\right) - \rho_1 + \frac{\rho_2}{2} + 0.5\right]$$

$$+\lambda^2\sigma^2\left[\frac{1}{2N^2}\left(2\sum_{i=1}^{N-1}(N-i)\rho_i + \sum_{i=2}^{N+1}(i-1)\rho_i + \sum_{i=N+2}^{2N}(2N+1-i)\rho_i + N\right) + \frac{\rho_2}{2} + \frac{1}{2} - \frac{1}{N}\left(\sum_{i=1}^{N-1}\rho_i + \sum_{i=2}^{N+1}\rho_i + 1\right)\right]$$

A Monte Carlo confirmation of the abovementioned relationship between MSE and lag-1 autocorrelation is illustrated in Figure S3.

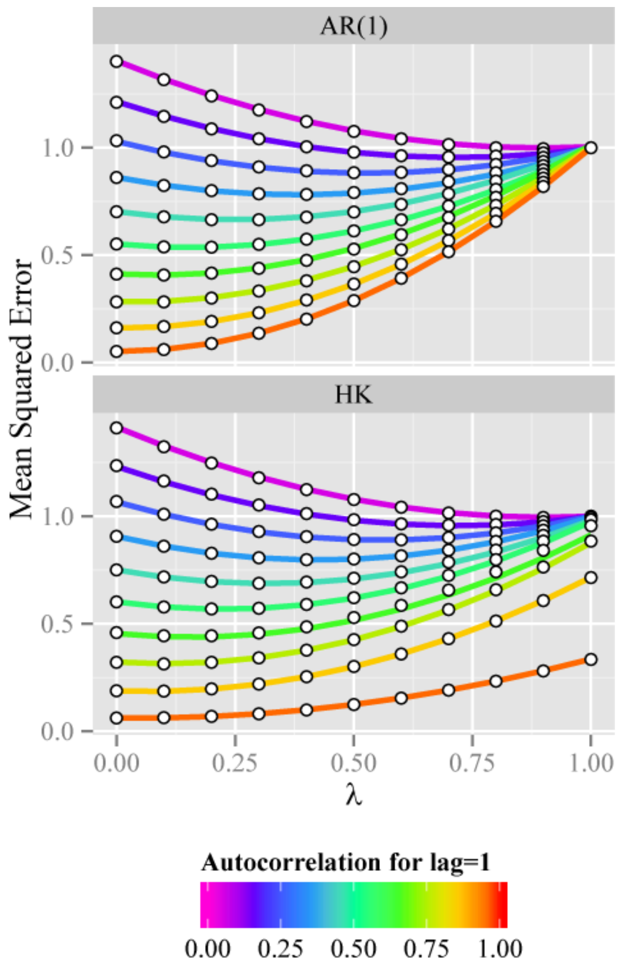

Figure S3. Monte Carlo confirmation of Eq. (8). Solid lines represent the Mean Squared Error (MSE) as estimated by Eq. (8) for different values of parameter λ, while the points correspond to the calculated MSE from the Monte Carlo simulations. Time series with 100000 values were generated from AR(1) and HK processes with zero mean and standard deviation equal to one and various values of lag-1 autocorrelation coefficient. The time series with HK dynamics were simulated using the function *SimulateFGN* from the R package FGN (Veenstra & McLeod, 2012).

4. Weighted Sum of local and total Average (WSA): method's sensitivity to time series length

Since the conclusions of the WSA methodology may depend on the overall length of the available time series (i.e., the term $2N$ in Eq. (8), where N is the number of available observations before or after the missing value), it is important to investigate the sensitivity of the presented methodology to different values of time series length (Figure S4).

As it is clearly illustrated in Figure S4, for the AR(1) process there is no significant effect on the minimum MSE vs λ relationship with the time series length. For processes presenting HK behavior (particularly for very high values of lag-1 autocorrelation), the optimal value of the parameter λ (i.e., the one that minimizes the MSE) depends strongly on the time series length. This is due to the nature of the HK processes. More specifically, when the available time series length is relatively small, the estimated global average is in essence a local rather than a global average. This peculiarity is therefore reflected in the optimal values of parameter λ, especially for high values of lag-1 autocorrelations (Figure S4). More specifically, since the parameter λ is the weighted factor ascribed to the overall global average (see also Eq. (7) in the main text), it should be expected that as the lag-1 autocorrelation increases, the value of λ that minimizes the MSE should also smoothly approach zero. This is indeed the case when the overall time series length is relatively high (Figure S4), but for shorter time series, given that what we estimate as global is rather a local average, the change of λ with the minimum MSE is more abrupt for high values of lag-1 autocorrelation (Figure S4).

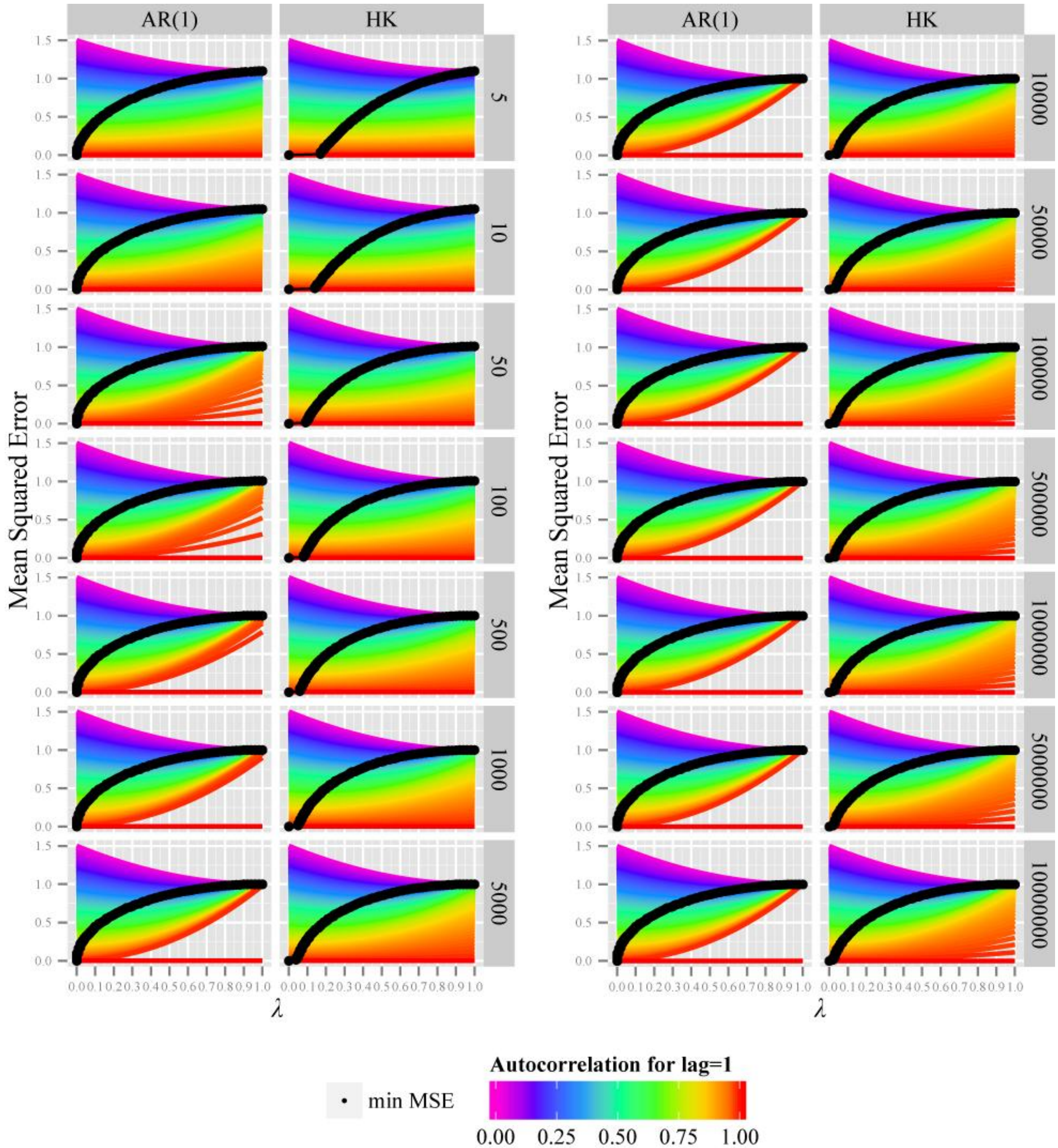

Figure S4. Sensitivity of the Mean Squared Error (MSE) estimation based on the Weighted Sum of local and total Average (WSA) method to the total time series length. The matrix of plots illustrates the relationship of MSE with the parameter λ for different values of lag-1 autocorrelation. The columns contain the results for processes with exponential (AR(1)) and power-law (HK) autocorrelation structure, while the rows include hypothetical time series lengths (from 2×5 to 2×10^7). While for AR(1) process there is no significant difference in the minimum MSE vs λ relationship with the time series length, for processes presenting HK behavior (particularly for very high values of lag-1 autocorrelation), the optimal value of the parameter λ (i.e., the one that minimizes the MSE) depends strongly on the time series length.

5. Weighted Sum of local and total Average (WSA): parameterization of the time series length

Figure S5 and S6 summarize the results of the sensitivity analysis to the overall time series length and the fitted functions to mimic these responses.

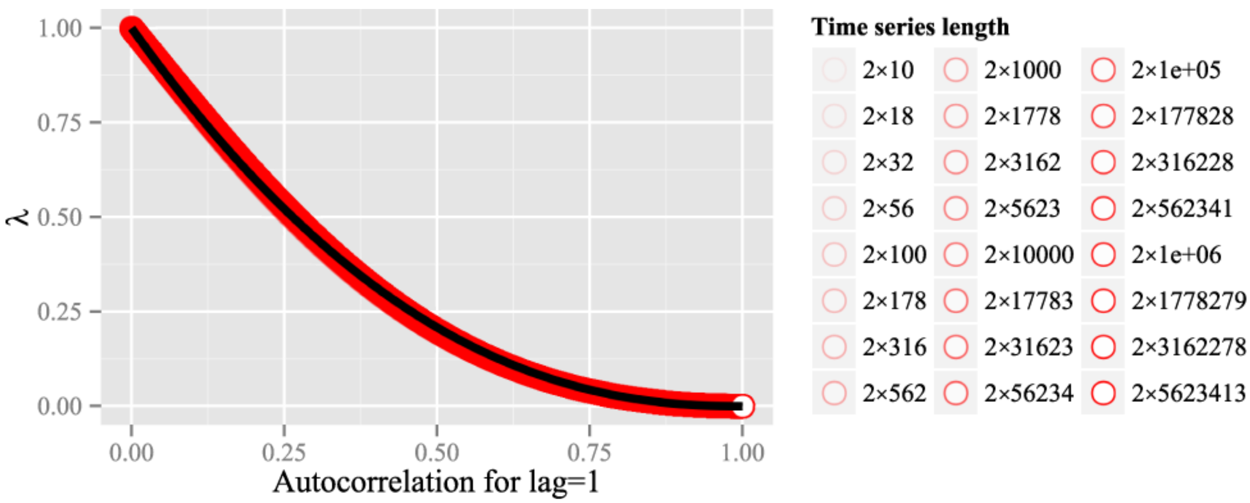

Figure S5. Optimal values (i.e., minimum MSE) of parameter λ, based on numerical experiment, for different lag-1 autocorrelations (ρ) and hypothetical time series lengths for processes with exponential (AR(1)) autocorrelation structure (red circles), as well as the fitted function describing the ρ vs λ relationship (Eq. (9) in the main text). There is no significant effect on the ρ vs λ relationship with the time series length.

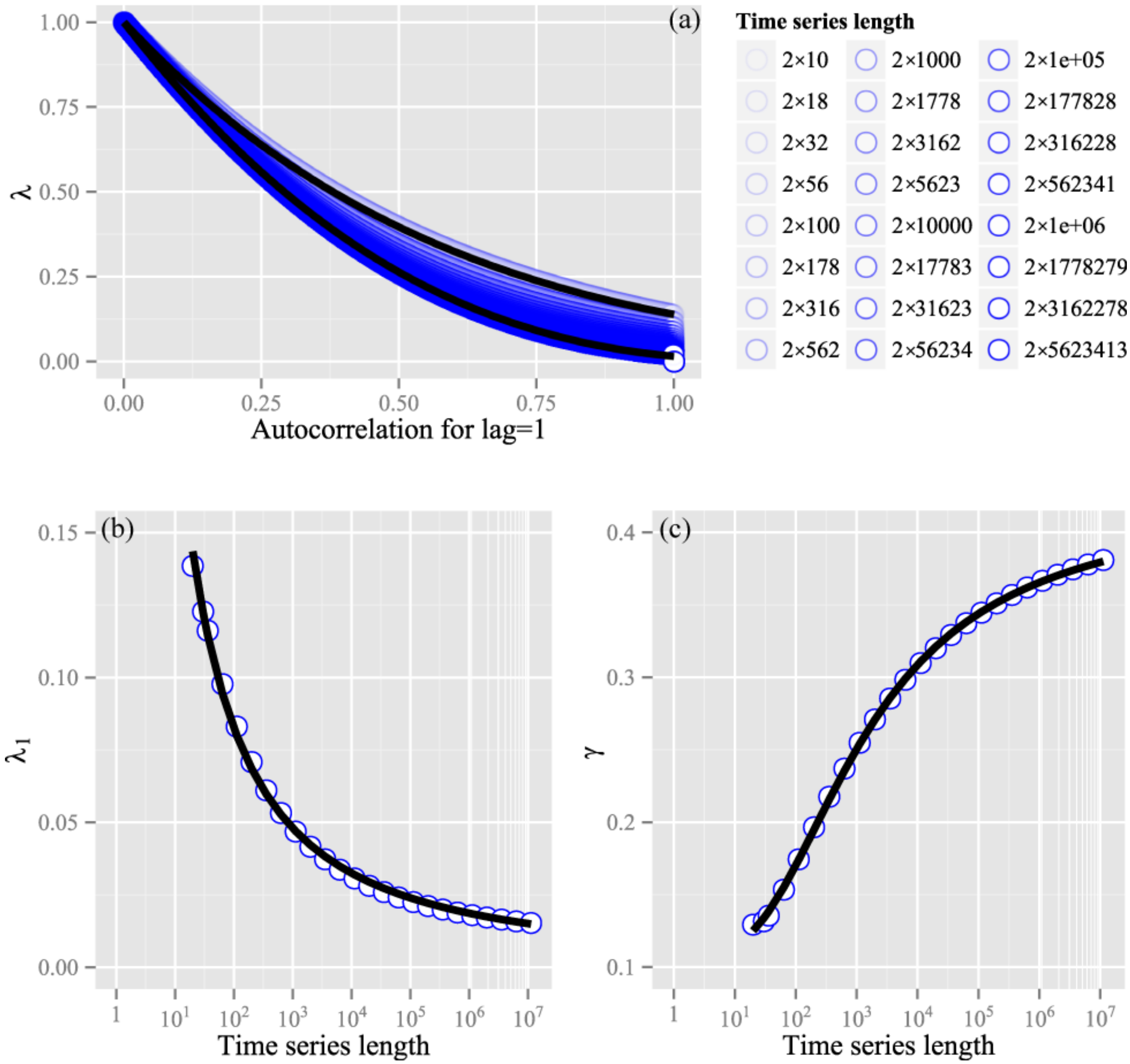

Figure S6. (a) Optimal values (i.e., minimum MSE) of parameter λ, based on numerical experiment, for different lag-1 autocorrelations (ρ) and hypothetical time series lengths for processes with power-law autocorrelation structure (blue circles), as well as the fitted function (solid black line; Eq. (10) in the main text). The optimal values of the parameter λ depend highly on the time series length. As the time series length increases, the ρ vs λ relationship of the HK process approaches the one of the AR(1). (b) Dependence of parameter λ_1 to the time series length. Parameter λ_1 reflects the value of parameter λ when $\rho \rightarrow 1$. (c) Dependence of parameter γ to the time series length. Blue circles correspond to the results of numerical experiment and black lines is the fitted function (λ_1 and γ are described n Eq. (10) of the main text).

6. References

Veenstra, J., & McLeod, A. I. (2012). Hyperbolic Decay Time Series Models. *In press*.