A quick gap filling of missing hydrometeorological data*

Christoforos Pappas[1,2], Simon Michael Papalexiou[3], and Demetris Koutsoyiannis[3]

[1] Département de géographie, Université de Montréal, Canada (**christoforos.pappas@umontreal.ca**)
[2] Institute of Environmental Engineering, ETH Zurich, Switzerland
[3] Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Greece

1. Introduction

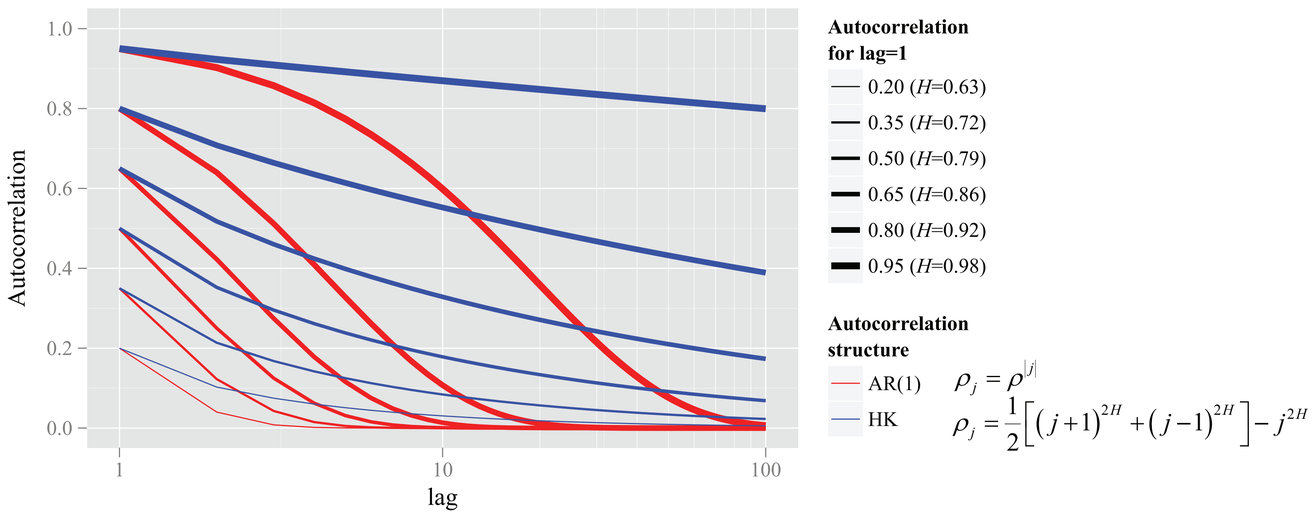

Figure 1. Theoretical autocorrelation functions for (i) Markovian processes, AR(1), with exponential decay of autocorrelation with lag and (ii) processes with HK behavior, described by the Hurst exponent *H*, with a power law relationship of autocorrelation with lag. The lag-1 autocorrelation, ρ, characterizes the strength of short-term persistence while the Hurst exponent, *H*, quantifies long-term dependences.

MOTIVATION AND RESEARCH GOAL:

Data gaps are ubiquitous in hydrometeorological time series, and filling these values still remains a challenge. Here, we present a quick and efficient gap-filling methodology for filling sporadic gaps based on the information content of the autocorrelation structure of the data.

KEY FINDINGS:

1. A definitive argument against the use of the sample average for filling correlated hydrometeorological data.

2. An innovative methodology, tailored for a quick filling of sporadic gaps, using information from time-adjacent values.

*Pappas, C., S. M. Papalexiou, and D. Koutsoyiannis (2014), *A quick gap filling of missing hydrometeorological data*, J. Geophys. Res. Atmos., 119, 9290–9300, doi:10.1002/2014JD021633.

2. Methodology

Optimal Local Average (OLA)

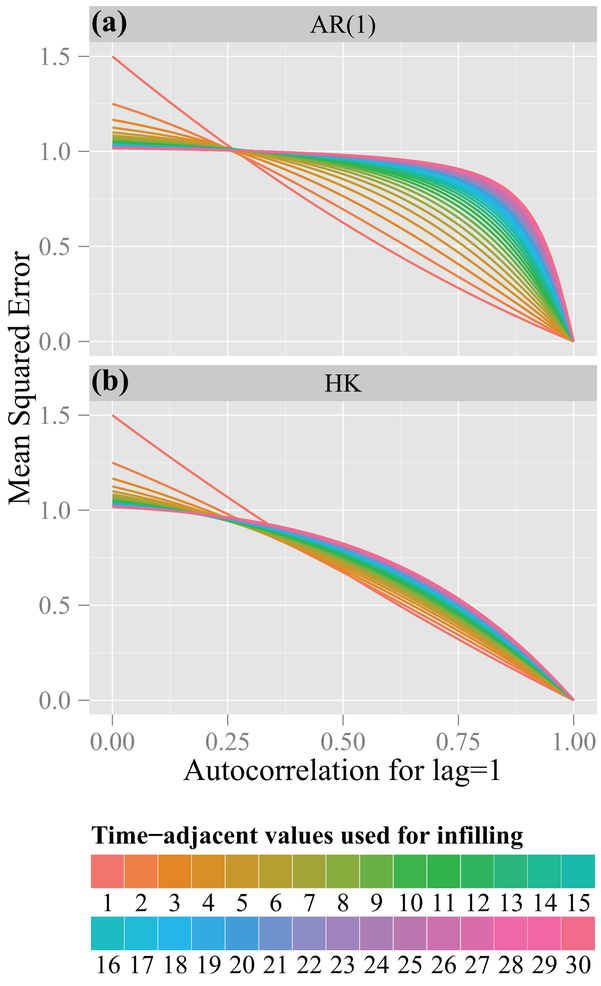

Figure 2. Illustration of the rationale underlying the Optimal Local Average (OLA) methodology for processes (a) with exponential and (b) with power law autocorrelation structure. The mean-squared error of an estimated missing value, based on local averages with different range (i.e., different number of neighboring values), for hypothetical time series with different lag-1 autocorrelation and standard deviation equal to 1, is depicted. When the number of time adjacent values used for the local average estimation equals 1, one value before and one after the missing observation are used for estimating the missing value, while when this number equals 30, the sample average is used (i.e., the average of all available observations, here for illustration assumed to be 30 before and 30 after the missing value).

$$\underline{y} = \frac{\sum_{i=1}^{n} x_{-i} + \sum_{i=1}^{n} x_i}{2n}$$

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

Weighted Sum of local and total Average (WSA)

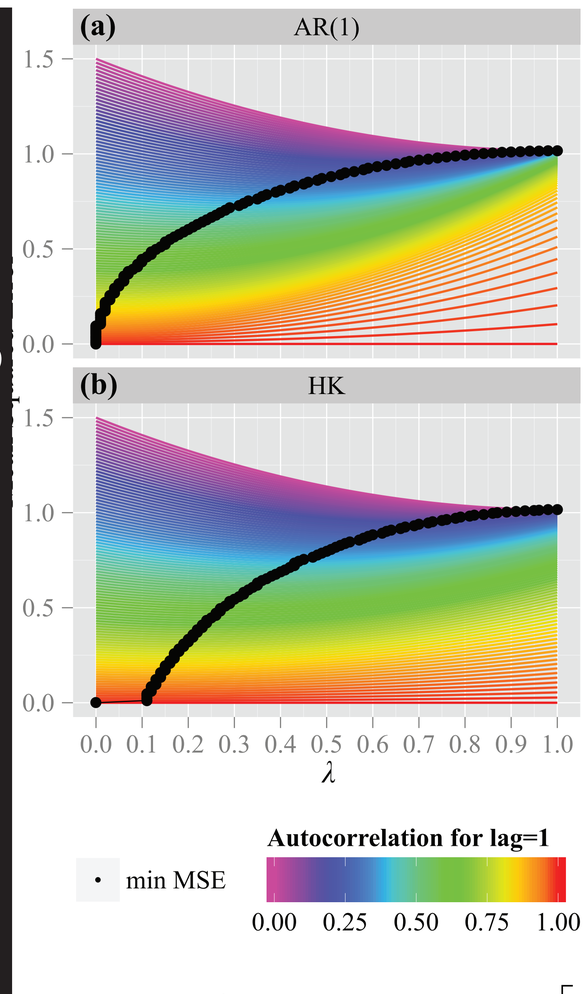

Figure 3. Surface plots of the mean-squared error (MSE) estimated according to the weighted sum of local and total average (WSA) methodology for different values of parameter λ, for processes with (a) exponential and (b) power law autocorrelation structure. The optimal values of parameter λ, i.e., the ones that minimize the MSE are also highlighted (black dots). As the lag-1 autocorrelation increases, the optimal value of parameter λ, which indicates the overall contribution of the global average, decreases.

$$\underline{y} = \lambda \frac{\sum_{i=-N}^{N} x_i}{2N} + (1-\lambda)\frac{x_{-1} + x_1}{2}$$

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

3. Case-studies

Monte Carlo simulations

Figure 4 (*left-hand side*). Estimated mean-squared error (MSE) based on different infilling methodologies (sample average i.e., using all the available values (here for illustration purposes 2 × 30 values are used); strictly local average using one observation before and one after the missing record; OLA; and WSA. Results correspond to processes with (a) exponential and (b) power law autocorrelation structure for different values of lag-1 autocorrelation. The solid lines depict the theoretical values of MSE while the dashed lines and uncertainty bounds correspond to the ensemble of the Monte Carlo simulations, filling artificial data gaps.

Figure 5 (*right-hand side*). Real-world examples of time series with Markovian behavior ((a) AR(1); annual precipitation) and with HK dynamics ((c) annual temperature and (e) annual minimum water depth). Original data are depicted in white circles, while the infilled time series are depicted in continuous colored lines.

Real World examples