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

Abstract

Data-gaps are ubiquitous in hydrometeorological time series and filling these values remains still a challenge. Since datasets without missing values may be a prerequisite in performing many analyses, a quick and efficient gap-filling methodology is required. In this study the problem of filling sporadic, single-value gaps using time-adjacent observations from the same location is investigated. The applicability of a local average (i.e., based on few neighboring in time observations) is examined and its advantages over the sample average (i.e., using the whole dataset) are illustrated. The analysis reveals that a quick and very efficient (i.e., minimum mean squared estimation error) gap-filling is achieved by combining a strictly local average (i.e., using one observation before and one after the missing value) with the sample mean.

Keywords: hydrometeorological data, missing values, gap-filling, interpolation, time series analysis

1. Introduction

Observing natural phenomena is of ultimate importance for understanding their complex characteristics. Understanding and simulating the earth-system processes requires a dense monitoring network of, not only long and reliable records, but also serially complete observations [e.g., *Butler*, 2014; *Baldocchi et al.*, 2012; *Silberstein*, 2006]. Time series of different geophysical variables (e.g., precipitation) emerge therefore from the systematic monitoring of their temporal evolution.

These instrumental time series are often plagued with a percentage of missing values (caused for example by malfunctioning of the equipment) creating sporadic and/or continuous gaps in their regular time-step. Many practical applications (e.g., extreme value analysis, continuous hydrological modeling) as well as statistical methodologies (e.g., spectral analysis, calibration (learning) algorithms, stochastic modeling and downscaling) have no tolerance to missing values. Preprocessing of raw datasets by infilling their missing values is thus a necessary procedure. Several interpolation techniques have been developed ranging from rather simple to extremely complex approaches. For example, *Henn et al.*, [2013], *Graham* [2009], *Horton and Kleinman* [2007], *Allison* [2003], *Roth* [1994], *Kemp et al.* [1983] provide detailed reviews of several gap-filling approaches applied to various scientific disciplines.

Several methods have been proposed for gap-filling environmental datasets e.g., linear or logistic regression, polynomial or spline interpolation, inverse distance weighting, ordinary kriging, and stochastic models that are fitted to the available records. More details on the aforementioned methodologies can be found in *Koutsoyiannis and Langousis* [2011] as well as in *Maidment*, [1993; ch. 19.4]. Additional statistical techniques that have been developed in the last decade, include artificial neutral networks and nearest neighbor techniques [*Elshorbagy et al.*, 2000, 2002], as well as approaches based on Kalman filter [*Alavi et al.*, 2006] and nonlinear mathematical programming [*Teegavarapu*, 2012]. Hybrid methods (employing both process-based and statistical tools) have been often also applied as part of weather generators [e.g., the MicroMet meteorological model; *Liston and Elder*, 2006]. Yet, the complexity and the computational demand of such methodologies often hamper their applicability to real world applications. While data-gaps are ubiquitous in hydrometeorological time series, how these gaps were filled is not often reported, or naïve approaches have been unjustifiably selected (e.g., such as filling the gaps with a fixed value, often corresponding to the sample average).

In this study, we present a definitive argument against the use of the sample average for filling correlated hydrometeorological data. In addition, an innovative methodology, tailored for a quick filling of sporadic (i.e., single-value) gaps using information from time-adjacent values of the same location (i.e., within-station method), is presented and its advantages over other commonly used approaches are illustrated. The present study provides therefore a quick gap-filling with high efficacy and is geared towards practitioners and data analysts.

2. Autocorrelation structure

Filling missing data, irrespective of the implemented statistical technique, requires a good understanding of the underlying process and its peculiarities. Although many properties are necessary for a complete description of the observed variables, their autocorrelation structure is of great importance. Autocorrelation describes the linear dependences among different values of a time series providing therefore insights on how their persistence evolves in time. As such, it is a key component in distilling information on the missing data and thus a cornerstone for the presented methodology.

Numerous studies illustrate different correlation patterns appropriate for describing several geophysical phenomena. Trying to cover the entire spectrum of autocorrelation structures widely detected and used in the hydrometeorological literature, we are focusing on: (i) processes with short-term persistence, characterized by exponential autocorrelation structure, and (ii) processes with long-term persistence, described by a power law autocorrelation function. These two structures have totally different characteristics in terms of time-dependence of the process and they are commonly present in different hydrometeorological variables (e.g., runoff, *Koutsoyiannis* [2013]; precipitation, *Marani* [2003]; sea level pressure and temperature, *Percival et al.* [2001], *Stephenson et al.* [2000]). Note that the suggested methodology is not limited to these two particular correlation structures. On the contrary, as it is demonstrated in the following sections, its applicability is more general, providing that the lag-1 autocorrelation can be estimated. Thus, the selection of known autocorrelation structures serves only for illustration of the theoretical framework underlying the methodology. In real-world applications, the estimation of empirical autocorrelations is enough for assuring the applicability and the efficacy of the proposed methodology.

In the following sections, hydrometeorological variables are treated as random variables modeled as stationary (more specifically, weakly stationary, i.e., with constant expected value and autocorrelation that depends only on the time lag; *Papoulis*, 1965, p.302) stochastic processes in discrete time. Regarding the notation used, the so-called Dutch notational convention is applied:

matrices and vectors are denoted by bold, random variables and stochastic processes are underlined, whereas their realizations (e.g. observed values) and the regular variables, are not.

2.1 Exponential autocorrelation structure

It has been very often claimed that hydrometeorological variables exhibit short-range dependence. For example, several studies have asserted that daily precipitation [*Gilman*, 1963], sea-surface temperature anomalies [*Frankignoul and Hasselmann*, 1977], Arctic sea ice [*Blanchard-Wrigglesworth et al.*, 2011; but see *Agarwal et al.*, 2012], climate variability [*Hasselmann*, 1976], as well as teleconnection patterns (such as North Atlantic Oscillation, Pacific-North American and West Pacific) [*Feldstein*, 2000; *Wunsch*, 1999; but see *Percival et al.*, 2001; *Stephenson et al.*, 2000] are characterized by Markovian dependence structure, i.e., the future appears to be independent of the past under the condition of known present [*Papoulis*, 1965, p.535]. This dependence is theoretically justified in a few cases, but appears to be physically implausible [*Koutsoyiannis and Montanari*, 2007; *Koutsoyiannis*, 2011]. Thus, the wide use of the model can be attributed to its simplicity rather than to its sound physical basis.

The Markovian property is reproduced by an autoregressive model of order one, AR(1), and the autocorrelation for different values of time lag j is given by:

$$\rho_j = \mathrm{corr}\left[\underline{x}_i, \underline{x}_{i+j}\right] = \frac{\mathrm{cov}\left[\underline{x}_i, \underline{x}_{i+j}\right]}{\mathrm{var}\left[\underline{x}_i\right]} = \rho^{|j|} \tag{1}$$

where ρ is the lag-1 autocorrelation coefficient $(|\rho|<1)$ which quantifies the short-range dependence. This relationship implies that the time dependence decreases exponentially as the time step (lag) increases, leading to practically negligible values of autocorrelation even for small values of lag (Figure 1).

2.2 Power law autocorrelation structure

There is strong empirical evidence that many natural phenomena are better characterized by highly persistent serial correlations rather than exponentially decaying autocorrelation structures. This natural behavior is often referred as Hurst phenomenon, long-term persistence, long-range dependence, or Hurst-Kolmogorov (HK) behavior [*Koutsoyiannis and Cohn*, 2008]. Here, the latter term is adopted, acknowledging the pioneering contribution of both, H. E. Hurst who first detected empirically that Nile river-level data exhibit long-term persistence [*Hurst*, 1951], and A. N. Kolmogorov who developed a basic mathematical framework describing this behavior [*Kolmogorov*, 1940].

HK behavior is identified in many diverse geophysical quantities such as wind power [*Bakker and van den Hurk*, 2012; *Haslett and Raftery*, 1989]; precipitation [*Fatichi et al.*, 2012; *Koutsoyiannis and Langousis*, 2011; *Montanari et al.*, 1996; *Savina et al.*, 2011, but see *Bunde et al.*, 2013]; snow depth [*Egli and Jonas*, 2009]; temperature [*Bloomfield*, 1992; *Gil-Alana*, 2005; *Scafetta and West*, 2005]; river discharge [Nile, Africa, *Koutsoyiannis*, 2002; Warta, Poland, *Radziejewski and Kundzewicz*, 1997; Po, Italy, *Montanari*, 2012; Tiber, Italy *Grimaldi*, 2004; Boeotikos Kephisos, Greece, *Koutsoyiannis*, 2003]; indices of North Atlantic Oscillation [*Stephenson et al.*, 2000]; solar activity [*Ogurtsov*, 2004; *Scafetta and West*, 2005; but see *Rypdal and Rypdal*, 2012]; extratropical atmospheric circulation anomalies [*Tsonis et al.*, 1999]; paleoclimate records [*Huybers and Curry*, 2006; *Markonis and Koutsoyiannis*, 2012].

A power law representation of autocorrelation decay with lag appears to be more appropriate for describing the temporal dependences of these phenomena. While lag-1 autocorrelation measures short-term persistence, the Hurst exponent, H $(0.5 < H < 1)$ is used to characterize the strength of the HK behavior (i.e., long-term persistence). For the case of random noise $H = 0.5$, whereas for real-world time series, like the examples mentioned above, H is often much higher. The autocorrelation function for lag *j*, is given by:

$$\rho_j = \frac{\operatorname{cov}\left[\underline{x}_i, \underline{x}_{i+j}\right]}{\operatorname{var}\left[\underline{x}_i\right]} = \frac{1}{2}\left(\left|j+1\right|^{2H} + \left|j-1\right|^{2H}\right) - \left|j\right|^{2H} \tag{2}$$

which for large j is proportional to j^{2H-2}. This behavior implies that the autocorrelation decreases according to a power-type function of lag, which is much slower than the exponential decay described by the Markovian dependence. Indeed, while the time dependence of the AR(1) process practically (for $\rho < 0.75$) vanishes for lag≈10, the autocorrelations of HK process are fat-tailed, maintaining significant values even for lags orders of magnitude higher (Figure 1).

3. Filling methods

Gap-filling techniques can be often presented as weighted averages of existing observations [*Koutsoyiannis and Langousis*, 2011] which can be summarized as follows:

$$\underline{y} = \mathbf{w}^T \underline{\mathbf{X}} + \underline{e} \tag{3}$$

where $\underline{y}$ is the missing value under examination $(\underline{y} \equiv \underline{x}_0)$, $\underline{\mathbf{X}} = \left[\underline{x}_{-N}, \ldots, \underline{x}_{-2}, \underline{x}_{-1}, \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\right]^T$ is a vector with the 2*N* random variables corresponding to the available observations (*T* denotes the transpose of the vector), $\mathbf{w} = \left[w_{-N}, \ldots, w_{-2}, w_{-1}, w_1, w_2, \ldots, w_N\right]^T$ is a vector with the weights assigned to each of the available observed values $\underline{\mathbf{X}}$, and $\underline{e}$ is the estimation error. Different infilling techniques provide therefore means for estimating the weighting parameters **w**. In the following sections, the Mean Squared Error (defined as $\mathrm{MSE} := \mathrm{E}\left[\underline{e}^2\right] = \sigma_e^2 + \mu_e^2$) is used as a performance metric for assessing different gap-filling approaches.

3.1 Optimal Local Average (OLA)

We assume that all the records used for the gap-filling ($x_{-n}, \ldots, x_{-2}, x_{-1}, x_1, x_2, \ldots, x_n$) have equal weights (i.e. $w_{-n} = \ldots = w_{-2} = w_{-1} = w_1 = w_2 = \ldots = w_n = 1/2n$; where $2n$ is the number of time-adjacent values used for estimating the missing value under examination). The estimated missing value under examination can then be expressed as $\underline{\hat{x}}_t = \left(\sum_{i=1}^{n} \underline{x}_{t-i} + \sum_{i=1}^{n} \underline{x}_{t+i}\right)/2n$ (i.e., arithmetic mean approach). Within the framework of Optimal Local Average (OLA) approach we test which is the optimal number of neighboring records (n) that should be used in order to have the best estimation of the missing value (i.e., the one that minimizes MSE). The MSE is given by:

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

Assuming that the underlying stochastic process is (weakly) stationary, we can express the MSE as a function of the standard deviation σ of the process, the number of the neighboring values n used for the infilling, and the correlation coefficient ρ_i for different values of lag i:

$$\mathrm{MSE} = \mathrm{E}\left[\underline{e}^2\right] = \frac{1}{2}\left(\frac{\sigma}{n}\right)^2 \left[(2n+1)\left(n - 2\sum_{i=1}^{n} \rho_i\right) + \sum_{i=1}^{2n} (2n+1-i)\rho_i\right] \tag{5}$$

The necessary algebraic manipulations for the derivation of Eq. (5) are detailed in the auxiliary material (S1). The resulting MSE for the two examined autocorrelation structures and different values of lag-1 autocorrelation (based on Eq. (5)) is illustrated in Figure 2 (see also Figure S2).

When processes with exponential autocorrelation structure are analyzed (Figure 2a and S2a), the strictly local average (i.e., $n = 1$) provides the minimum MSE for a wide range of lag-1 autocorrelations. There is a critical value of lag-1 autocorrelation ($\rho_{\mathrm{cr}}^{\mathrm{AR}} = 0.29$) above which the strictly local average provides the best estimate (Table 1 and Figure 2a and S2a). This manifests the fundamental Markovian property underlying the AR(1) process, i.e., when ρ becomes non-negligible ($\rho \geq 0.29$) the information content of neighboring values should only be used, otherwise the MSE is larger (Figure 2a and S2a). For the case of power law autocorrelation structure, the time-adjacent values required for a minimum MSE decrease gradually as the lag-1 autocorrelation increases (Figure 2b and S2b), as opposed to the sharp response of the AR(1) processes (Figure 2a and S2a). A critical value of ρ above which the strictly local average is preferable still exist but it is higher than the one of AR(1) processes ($\rho_{\mathrm{cr}}^{\mathrm{HK}} = 0.52$; Table 1 and Figure 2b and S2b).

In summary, no matter which is the underlying autocorrelation structure (exponential or power-type) when $\rho \geq 0.52$ the strictly local average (i.e., using one observation before and one after the missing value) provides the best estimate. Moreover, as Figure 2 and S2 illustrate, for a wide range of lag-1 autocorrelations, the sample average inflates the MSE, and therefore should be avoided when correlated data have to be infilled.

3.2 Weighted Sum of local and total Average (WSA)

Building upon the aforementioned findings (i.e., the time-adjacent values used for an efficient gap-filling can be determined by the lag-1 autocorrelation of the examined data), we provide a generalized framework distilling information from both, local and sample (global) average. The sum of the strictly local (i.e., one value before and one after the missing record) and sample average, weighted according to the lag-1 autocorrelation of the examined data, is used as an estimation of the missing value. The estimated missing value is then given by:

$$\underline{\hat{x}}_t = \lambda \frac{\sum_{i=1}^{N} \left(\underline{x}_{t-i} + \underline{x}_{t+i} \right)}{2N} + (1-\lambda) \frac{\underline{x}_{t-1} + \underline{x}_{t+1}}{2} \tag{6}$$

where λ is the weighting factor assigned to the sample average. In essence, parameter λ reflects the strength of the temporal correlation i.e., low (high) values of λ imply low (high) contribution of the sample average and thus high (low) temporal autocorrelation. Under the assumption of (weak) stationarity, the MSE is given by:

$$\mathrm{MSE} = \mathrm{E}\left[\underline{e}^2 \right] = \mathrm{E}\left[\left(\underline{x}_t - \underline{\hat{x}}_t \right)^2 \right] = \mathrm{E}\left[\left(\underline{x}_t - \left(\lambda \frac{\sum_{i=1}^{N} \left(\underline{x}_{t-i} + \underline{x}_{t+i} \right)}{2N} + (1-\lambda) \frac{\underline{x}_{t-1} + \underline{x}_{t+1}}{2} \right) \right)^2 \right] \tag{7}$$

where $2N$ is the length of the available observations. After some algebraic manipulations, detailed in the auxiliary material (S3), the following expression is obtained:

$$\begin{aligned} \mathrm{MSE} = {} & \frac{1}{2}\sigma^2 (3 - 4\rho_1 + \rho_2) - 2\lambda\sigma^2 \left[\frac{1}{N} \sum_{i=1}^{N} \rho_i - \frac{1}{2N} \left(\sum_{i=1}^{N-1} \rho_i - \sum_{i=2}^{N+1} \rho_i + 1 \right) - \rho_1 + \frac{\rho_2}{2} + 0.5 \right] \\ & + \lambda^2 \sigma^2 \left[\begin{aligned} & \frac{1}{2N^2} \left(2 \sum_{i=1}^{N-1} (N-i) \rho_i + \sum_{i=2}^{N+1} (i-1) \rho_i + \sum_{i=N+2}^{2N} (2N+1-i) \rho_i + N \right) \\ & + \frac{\rho_2}{2} + \frac{1}{2} - \frac{1}{N} \left(\sum_{i=1}^{N-1} \rho_i + \sum_{i=2}^{N+1} \rho_i + 1 \right) \end{aligned} \right] \end{aligned} \tag{8}$$

The MSE is therefore expressed as a function of σ, ρ_i, N, and the weighting factor λ. For a given autocorrelation structure (i.e., known ρ_i), we seek the value of λ that yields the minimum MSE (λ_{opt}; black dots in Figure 3; see also Figure S5 and S6a). For both Markovian and HK behavior, as lag-1 autocorrelation increases the contribution of the local average increases (i.e., λ_{opt} decreases, Figure 3; see also Figure S5 and S6a).

An interesting property of the HK behavior is reflected in Figure 3b (see also Figure S6a): for high values of lag-1 autocorrelation there is a discontinuity in the values of λ_{opt}. More specifically, while it is expected that for high lag-1 autocorrelation the sample average does not contribute at all to the estimation of the missing value (i.e., $\lambda_{opt} = 0$), λ does not reach zero gradually and has non-zero values even for high values of lag-1 autocorrelation (Figure 3b and S6a). The rationale behind this behavior is that it takes time for a process with long-range dependence to reveal its characteristics. More specifically, when the available time series length is relatively small, the estimated sample average is in essence a local rather than a global average (see also detailed discussion in the auxiliary material (S4)).

In order to assess the influence of sample size in the MSE estimation and thus in λ_{opt}, a sensitivity analysis was conducted checking sample sizes from 2×5 to 2×10^7 (auxiliary material S4 and S5). For processes with exponential autocorrelation structure the relationship of λ_{opt} with ρ does not vary much with the time series length (Figure S4 and S5) and it is thus approximated by:

$$\lambda_{\mathrm{opt}}^{\mathrm{AR(1)}} = (1-\rho)^{2.26} \tag{9}$$

For processes with HK behavior λ_{opt} depends highly on the time series length (Figure S4 and S6). To mimic this type of dependence the λ_{opt} vs ρ relationship is approximated using two additional parameters (λ_1, γ):

$$\lambda_{\text{opt}}^{\text{HK}} = \left(1 - \left(1 - \lambda_1^{\gamma}\right)\rho\right)^{1/\gamma} \tag{10}$$

where $\lambda_1 = 0.70 / \left(1 + \ln^2\left(N\right)\right)^{0.69}$ and $\gamma = 0.44 - 0.33 / \left(1 + \ln^2\left(1 + 0.03\ln^2\left(N\right)\right)\right)$.

In practice, for Markovian processes, once the lag-1 autocorrelation is estimated from the data, it can be plugged-in to Eq. (9) and estimate the value of λ_{opt}, then, Eq. (6) can be applied for gap-filling the examined missing value. For data with HK behavior, one additional (but very simple) step is needed, i.e., the calculation of λ_1 and γ given the length of the available records.

4. Methods intercomparison and discussion

The presented methods (OLA and WSA) are in essence generalizations of the widely applied concept of arithmetic mean, enhanced with information from the lag-1 autocorrelation of the examined time series. Their performance is tested against the sample average and the strictly local average (i.e., a linear interpolation between the values adjacent to the missing value) by comparing the resulting estimation error (Figure 4 and Figure 5).

4.1 Monte-Carlo simulations

Synthetic time series with 100000 values were generated from AR(1) and HK processes with zero mean and standard deviation equal to one and with lag-1 autocorrelation coefficients covering the entire range of possible values. The time series with HK dynamics were simulated using the function SimulateFGN from the `R` package FGN [*Veenstra and McLeod*, 2012]. Each value of the time series was then sequentially removed and the artificial data gap was then filled with the OLA and WSA approach, as well as we the sample and the strictly local average.

For the entire range of lag-1 autocorrelation for both exponential and power-type autocorrelation structures, the WSA methodology provides the minimum MSE (Figure 4). When uncorrelated data are examined (i.e., lag-1 autocorrelation tends to zero) the WSA method converge to the sample average approach, while for highly correlated time series, it converges to the strictly local average approach. For relatively strongly correlated data (e.g., with lag-1 autocorrelation higher than 0.5) WSA and OLA methods have similar performance but the flexible character of WSA approach makes it appropriate for gap-filling sporadic gaps across the entire spectrum of temporal dependences (minimum MSE; Figure 4). Similarly to the inverse distance weighting (where information from neighboring in space station is used), in WSA approach the notion of similarity (in time) between data points is crucial. The WSA methodology is therefore based on Tobler's first law in geography i.e., "everything is related to everything else, but near things are more related than distant things" [*Tobler*, 1970].

The sample average provides the worst results (highest MSE; Figure 4). Replacing a missing value with the sample average is a simple and easy approach for dealing with missing data, but as our analysis reveals much better results can be obtained by applying tools of similar complexity (i.e., WSA approach). It is also worth mentioning that a great advantage of the WSA methodology is that the probability distribution of the observations and the temporal relationships (i.e., autocorrelation) remain relatively undisturbed, avoiding induction of biases in the mean, variance or autocorrelation of the final time series. As already underlined elsewhere [e.g., *Little and Rubin*, 2002], replacing all missing values in a dataset with a single value (e.g., using the sample mean) apart from reducing the variance, can often artificially inflate the significance of any statistical test that is based on these statistics. The proposed method is not free of these problems, but if the autocorrelation is strong and the percentage of missing values low, the reduction of variance is not substantial. In addition, with the WSA methodology no assumption was made regarding the distribution underlying the examined dataset. Moreover, as demonstrated in Eq. (8), there is no dependence of the MSE to the mean properties of the analyzed dataset, therefore there is no need for data preprocessing (e.g., normalization). Given that the underlying autocorrelation structure is identified, this simple method does not impose any requirement for calculation of other statistical quantities apart from the sample mean and the lag-1 autocorrelation for its application.

4.2 Real-world applications

An additional illustration of the aforementioned finding is summarized in Figure 5. Real-world time series from the Global Historical Climatology Network, (GHCN version 2.60; www.ncdc.noaa.gov/oa/climate/ghcn-daily) and from the Roda Nilometer, near Cairo (minimum water levels of Nile; [*Toussoun*, 1925]) are presented (Figure 5).

Time series of annual precipitation (GHCN, station ID: CA003031093) behaving as Markovian process with lag-1 autocorrelation $\rho = 0.29$, and temperature (GHCN, station ID: GM000003342) presenting HK dynamics with Hurst exponent $H = 0.72$ (estimated using the slope of the climacogram, which is a double-logarithmic plot of the standard deviation of the sample at an aggregate timescale vs the timescale; *Koutsoyiannis*, 2003, 2010) are examined. Observed records spanning from 1893 to 2011, were infilled removing sequentially every single value and gap-filling with the presented methodologies (Figure 5). The original and the infilled time series, as well as the efficiency of each infilling approach (defined as $\text{1-nRMSE} = 1 - \sqrt{\text{MSE}}/s$ where nRMSE is the normalized Root Mean Squared Error and s is the standard deviation of the observed time series) are presented in Figure 5a,c and Figure 5b,d respectively.

The longest instrumental record of the water levels of Nile is also examined (focusing on the period 622 AD to 1470 AD that the record is almost uninterrupted), illustrating the advantages of within-station gap-filling approaches. In this case, the use of within-station information is apparently the only solution since neighboring stations, covering the same time period, are not available. The annual minimum water levels of Nile are characterized by HK dynamics with Hurst exponent $H = 0.87$ [*Koutsoyiannis*, 2013]. For clarity in the illustration, only 200 years are presented, covering the period 800 AD to 1000 AD (Figure 5e,f), but the processing was made, and the efficiency was calculated, for the entire series. The WSA approach yields the highest efficiency (Figure 5f). Since the data present strong autocorrelation, gap-filling with OLA and local average approach converges to the same results (Figure 5e,f; see also Table 1). As in the previous two real-world examples (i.e., annual precipitation and temperature time series), the sample average has no skill (i.e., efficiency tends to 0; Figure 5b,d,f) manifesting that its use in gap-filling hydrometeorological variables is not only unjustified, but also seriously flawed.

In accordance with the presented theory (Section 3) and Monte-Carlo simulations (Figure 4), WSA approach provides the highest efficiency ($1 - \text{nRMSE}$; Figure 5b,d,f). It is worth underline that, when the examined time series do not present high autocorrelation, the other three approaches (i.e., sample and local average as well as OLA) lead to comparable results in terms of overall efficiency (Figure 5b,d), but the distribution of the infilled time series varies significantly (Figure 5a,c). When data with strong autocorrelation are examined (Figure 5f), the use of neighboring in time values improves significantly the performance of the gap-filling approach (Figure 5e,f). As expected, the use of sample average vanishes the variability presented in the original record, while the local average preserves many interesting features of the original records (Figure 5a,c,e).

5. Limitations and further improvements

While our analysis is focusing on sporadic, single-value data-gaps, generalizations of the presented approach for a wider gap-window are possible. Continuous missing values can be infilled by applying sequentially the presented framework (cascade process). More specifically, the available observations, at the gap-window boundaries, are used for the estimation of the missing value in the middle of the gap-window. This value is then used as a proxy, applying again the WSA approach for the new, restricted gap-window. A cascade-based procedure can be thus applied gap-filling sequentially continuous missing values. However, it is worth mentioning that the wider the data-gaps the more uncertain the estimated first- and higher-order statistics of the examined time series and thus the estimated missing values. A more elegant (but computationally more demanding) approach for dealing with multiple sequential data-gaps can emerge from Eq. (3). More specifically, instead of assigning equal weights to the available observations (as is the case for the presented methods; see Section 3), for each of the examined missing values, specific weighting factors ($\mathbf{w}$) can be assigned to the available observations by solving explicitly Eq. (3); results of ongoing research on this issue are planned to be presented soon.

6. Conclusions

Conventional methods for handling missing data (such as sample average or linear interpolation of values adjacent to the missing record) are seriously flawed in the hydrometeorological time series, where the time series autocorrelation is non-negligible. Taking advantage of the information content of the lag-1 autocorrelation, a new flexible and equally simple framework for a quick gap filling of sporadic, single-value, gaps is proposed. The conclusions of our study are twofold: (i) a definitive argument against the use of the sample average for infilling correlated data is provided and demonstrated theoretically; and (ii) a new gap-filling methodology, equivalently simple but significantly more efficient, using a weighted sum of sample and strictly local average, is developed and its advantages are illustrated. The estimation of the sample mean and the lag-1 autocorrelation is the only necessity for assuring the applicability of WSA approach. The presented methodology is therefore a valuable tool for a quick filling of a small number of missing measurements tailored for hydrometeorological data as well as for a efficient gap-filling of missing paleoclimatic records, where neighboring station are not available.

Acknowledgments

We would like to thank the eponymous reviewer Prof. Salvatore Grimaldi and two anonymous reviewers for their constructive comments that helped us improve the presentation of our analysis. Fruitful discussions with Y. Dialynas, P. Kossieris, K. Kyriakidis, A. Lykou, and Y. Markonis, at an early stage of this analysis, are also gratefully acknowledged. The computer code used for the presented methodologies is available upon request from the corresponding author. All statistical analyses and graphics were performed using the `R` environment [*R*

CoreTeam, 2012] as well as additional packages as *ggplot2* [*Wickham*, 2009] and *SimulateFGN* [*Veenstra and McLeod*, 2012]. The authors would like to thank the developers for making them freely available. Global Historical Climatology Network is also acknowledged for providing the presented time series.

References

Agarwal, S., W. Moon, and J. S. Wettlaufer (2012), Trends, noise and re-entrant long-term persistence in Arctic sea ice, *Proc. R. Soc. A Math. Phys. Eng. Sci.*, *468*(2144), 2416–2432, doi:10.1098/rspa.2011.0728.

Alavi, N., J. S. Warland, and A. a. Berg (2006), Filling gaps in evapotranspiration measurements for water budget studies: Evaluation of a Kalman filtering approach, *Agric. For. Meteorol.*, *141*(1), 57–66, doi:10.1016/j.agrformet.2006.09.011.

Allison, P. (2003), Missing data techniques for structural equation modeling, *J. Abnorm. Psychol.*, *112*(4), 545–557, doi:10.1037/0021-843X.112.4.545.

Bakker, A. M. R., and B. J. J. M. van den Hurk (2012), Estimation of persistence and trends in geostrophic wind speed for the assessment of wind energy yields in Northwest Europe, *Clim. Dyn.*, *39*(3-4), 767–782, doi:10.1007/s00382-011-1248-1.

Baldocchi, D., M. Reichstein, D. Papale, L. Koteen, R. Vargas, D. Agarwal, and R. Cook (2012), The Role of Trace Gas Flux Networks in the Biogeosciences, *EOS Trans.*, *93*(23), 217–224.

Blanchard-Wrigglesworth, E., K. C. Armour, C. M. Bitz, and E. DeWeaver (2011), Persistence and Inherent Predictability of Arctic Sea Ice in a GCM Ensemble and Observations, *J. Clim.*, *24*(1), 231–250, doi:10.1175/2010JCLI3775.1.

Bloomfield, P. (1992), Trends in global temperature, *Clim. Change*, *21*(1), 1–16, doi:10.1007/BF00143250.

Bunde, A., U. Büntgen, J. Ludescher, J. Luterbacher, and H. von Storch (2013), Is there memory in precipitation?, *Nat. Clim. Chang.*, *3*(3), 174–175, doi:10.1038/nclimate1830.

Butler, D. G. (2014), Earth observation enters next phase, *Nature*, *508*, 160–161.

Egli, L., and T. Jonas (2009), Hysteretic dynamics of seasonal snow depth distribution in the Swiss Alps, *Geophys. Res. Lett.*, *36*(2), L02501, doi:10.1029/2008GL035545.

Elshorbagy, A., U. Panu, and S. Simonovic (2000), Group-based estimation of missing hydrological data: I. Approach and general methodology, *Hydrol. Sci. J.*, *45*(6), 849–866.

Elshorbagy, A., S. Simonovic, and U. Panu (2002), Estimation of missing streamflow data using principles of chaos theory, *J. Hydrol.*, *255*, 123–133.

Fatichi, S., V. Y. Ivanov, and E. Caporali (2012), Investigating Interannual Variability of Precipitation at the Global Scale: Is There a Connection with Seasonality?, *J. Clim.*, *25*(16), 5512–5523, doi:10.1175/JCLI-D-11-00356.1.

Feldstein, S. B. (2000), The Timescale, Power Spectra, and Climate Noise Properties of Teleconnection Patterns, *J. Clim.*, *13*(24), 4430–4440, doi:10.1175/1520-0442(2000)013<4430:TTPSAC>2.0.CO;2.

Frankignoul, C., and K. Hasselmann (1977), Stochastic climate models, Part II Application to sea- • surface temperature anomalies and thermocline variability, *Tellus*, *29*, 289–305.

Gil-Alana, L. (2005), Statistical modeling of the temperatures in the Northern Hemisphere using fractional integration techniques, *J. Clim.*, *18*(24), 5357–5369.

Gilman, D. (1963), On the power spectrum of “red noise,” *J. Atmos. Sci.*, *20*, 182–184.

Graham, J. (2009), Missing data analysis: Making it work in the real world, *Annu. Rev. Psychol.*, *60*, 549–576, doi:10.1146/annurev.psych.58.110405.085530.

Grimaldi, S. (2004), Linear parametric models applied to daily hydrological series, *J. Hydrol. Eng.*, *9*, 383–391.

Haslett, J., and A. Raftery (1989), Space-time modelling with long-memory dependence: Assessing Ireland’s wind power resource, *J. R. Stat. Soc. Ser. C (Applied Stat.*, *38*(1), 1–50.

Hasselmann, K. (1976), Stochastic climate models Part I. Theory, *Tellus*, *28*(6), 473–485, doi:10.1111/j.2153-3490.1976.tb00696.x.

Henn, B., M. S. Raleigh, A. Fisher, and J. D. Lundquist (2013), A Comparison of Methods for Filling Gaps in Hourly Near-Surface Air Temperature Data, *J. Hydrometeorol.*, *14*(3), 929–945, doi:10.1175/JHM-D-12-027.1.

Horton, N. J., and K. P. Kleinman (2007), Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models, *Am. Stat.*, *61*(1), 79–90, doi:10.1198/000313007X172556.

Hurst, H. (1951), Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.*, (2447).

Huybers, P., and W. Curry (2006), Links between annual, Milankovitch and continuum temperature variability, *Nature*, *441*(7091), 329–32, doi:10.1038/nature04745.

Kemp, W., D. Burnell, D. Everson, and A. Thomson (1983), Estimating missing daily maximum and minimum temperatures, *J. Clim. Appl. Meteorol.*, *22*, 1587–1593.

Kolmogorov, A. (1940), Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen Raum, *Dokl. Acad. Sci. URSS*, *26*, 115–118.

Koutsoyiannis, D. (2002), The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrol. Sci. J.*, *47*(4), 37–41.

Koutsoyiannis, D. (2003), Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrol. Sci. J.*, *48*(1), 3–24, doi:10.1623/hysj.48.1.3.43481.

Koutsoyiannis, D. (2010), HESS Opinions “A random walk on water,” *Hydrol. Earth Syst. Sci.*, *14*(3), 585–601, doi:10.5194/hess-14-585-2010.

Koutsoyiannis, D. (2011), Hurst–Kolmogorov dynamics as a result of extremal entropy production, *Phys. A Stat. Mech. its Appl.*, *390*(8), 1424–1432, doi:10.1016/j.physa.2010.12.035.

Koutsoyiannis, D. (2013), Hydrology and Change, *Hydrol. Sci. J.*, doi:10.1080/02626667.2013.804626.

Koutsoyiannis, D., and T. Cohn (2008), The Hurst phenomenon and climate, *EGU Gen. Assem.*, *Session IS*(Climatic and hydrological perspectives on long-term changes).

Koutsoyiannis, D., and A. Langousis (2011), Precipitation, in *Treatise on Water Science*, edited by P. Wilderer and S. Uhlenbrook, pp. 27–78, Academic Press, Oxford.

Koutsoyiannis, D., and A. Montanari (2007), Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resour. Res.*, *43*(5), W05429, doi:10.1029/2006WR005592.

Liston, G., and K. Elder (2006), A meteorological distribution system for high-resolution terrestrial modeling (MicroMet), *J. Hydrometeorol.*, *7*, 217–234.

Little, R., and D. Rubin (2002), *Statistical analysis with missing data*, Second Edi., John Wiley and Sons, New York.

Maidment, D. (1993), *Handbook of hydrology*, McGraw-Hill, New York.

Marani, M. (2003), On the correlation structure of continuous and discrete point rainfall, *Water Resour. Res.*, *39*(5), 1128, doi:10.1029/2002WR001456.

Markonis, Y., and D. Koutsoyiannis (2012), Climatic Variability Over Time Scales Spanning Nine Orders of Magnitude: Connecting Milankovitch Cycles with Hurst-Kolmogorov Dynamics, *Surv. Geophys.*, doi:10.1007/s10712-012-9208-9.

Montanari, A. (2012), Hydrology of the Po River: looking for changing patterns in river discharge, *Hydrol. Earth Syst. Sci.*, *16*(10), 3739–3747, doi:10.5194/hess-16-3739-2012.

Montanari, A., R. Rosso, and M. Taqqu (1996), Some long-run properties of rainfall records in Italy, *J. Geophys. Res.*, *101*(D23), 29431–29438.

Ogurtsov, M. (2004), New Evidence for Long-Term Persistence in the Sun's Activity, *Sol. Phys.*, *220*, 93–105.

Papoulis, A. (1965), *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill.

Percival, D., J. Overland, and H. Mofjeld (2001), Interpretation of North Pacific Variability as a Short-and Long-Memory Process, *J. Clim.*, *14*, 4545–4559.

Radziejewski, M., and Z. Kundzewicz (1997), Fractal analysis of flow of the river Warta, *J. Hydrol.*, *200*, 280–294.

RCoreTeam (2012), *R: A Language and Environment for Statistical Computing*, Vienna, Austria.

Roth, P. L. (1994), Missing Data: a Conceptual Review for Applied Psychologists, *Pers. Psychol.*, *47*(3), 537–560, doi:10.1111/j.1744-6570.1994.tb01736.x.

Rypdal, M., and K. Rypdal (2012), Is there long-range memory in solar activity on timescales shorter than the sunspot period?, *J. Geophys. Res.*, *117*(A4), A04103, doi:10.1029/2011JA017283.

Savina, M., P. Molnar, and P. Burlando (2011), Seasonal long-term persistence in radar precipitation in complex terrain, *Water Resour. Res.*, *47*(10), W10506, doi:10.1029/2010WR010170.

Scafetta, N., and B. J. West (2005), Multiscaling comparative analysis of time series and geophysical phenomena, *Complexity*, *10*(4), 51–56, doi:10.1002/cplx.20076.

Silberstein, R. P. (2006), Hydrological models are so good, do we still need data?, *Environ. Model. Softw.*, *21*(9), 1340–1352, doi:10.1016/j.envsoft.2005.04.019.

Stephenson, D. B., V. Pavan, and R. Bojariu (2000), Is the North Atlantic Oscillation a random walk?, *Int. J. Climatol.*, *20*(1), 1–18, doi:10.1002/(SICI)1097-0088(200001)20:1<1::AID-JOC456>3.0.CO;2-P.

Teegavarapu, R. (2012), Spatial interpolation using nonlinear mathematical programming models for estimation of missing precipitation records, *Hydrol. Sci. J.*, *57*(3), 37–41.

Tobler, W. (1970), A computer movie simulating urban growth in the Detroit region, *Econ. Geogr.*, *46*, 234–240.

Toussoun, O. (1925), *Mémoire sur l'histoire du Nil*.

Tsonis, A. A., P. J. Roebber, and J. B. Elsner (1999), Long-Range Correlations in the Extratropical Atmospheric Circulation: Origins and Implications, *J. Clim.*, *12*(5), 1534–1541, doi:10.1175/1520-0442(1999)012<1534:LRCITE>2.0.CO;2.

Veenstra, J., and A. I. McLeod (2012), Hyperbolic Decay Time Series Models, *Press*.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Springer New York.

Wunsch, C. (1999), The Interpretation of Short Climate Records, with Comments on the North Atlantic and Southern Oscillations, *Bull. Am. Meteorol. Soc.*, *80*(2), 245–255, doi:10.1175/1520-0477(1999)080<0245:TIOSCR>2.0.CO;2.

Tables

Table 1. Time-adjacent values needed for an optimal infilling (minimum Mean Squared Error) of missing observations according to the Optimal Local Average methodology, for different autocorrelation structures (short or long-term persistence) and the lag-1 autocorrelations.

Optimal Local Average			
Short-term persistence		Long-term persistence	
$\rho \leq 0.25$	$n = n_{\text{max}}$	$\rho < 0.3$	$n = n_{\text{max}}$
$0.26 \leq \rho \leq 0.28$	$n = 2$	$0.30 \leq \rho \leq 0.32$	$n = 4$
		$0.33 \leq \rho \leq 0.38$	$n = 3$
$\rho \geq 0.29$	$n = 1$	$0.39 \leq \rho \leq 0.51$	$n = 2$
		$\rho \geq 0.52$	$n = 1$

ρ : lag-one autocorrelation coefficient

n : time-adjacent values used for the infilling

n_{max} : all the available observed values, i.e., sample average

Figures

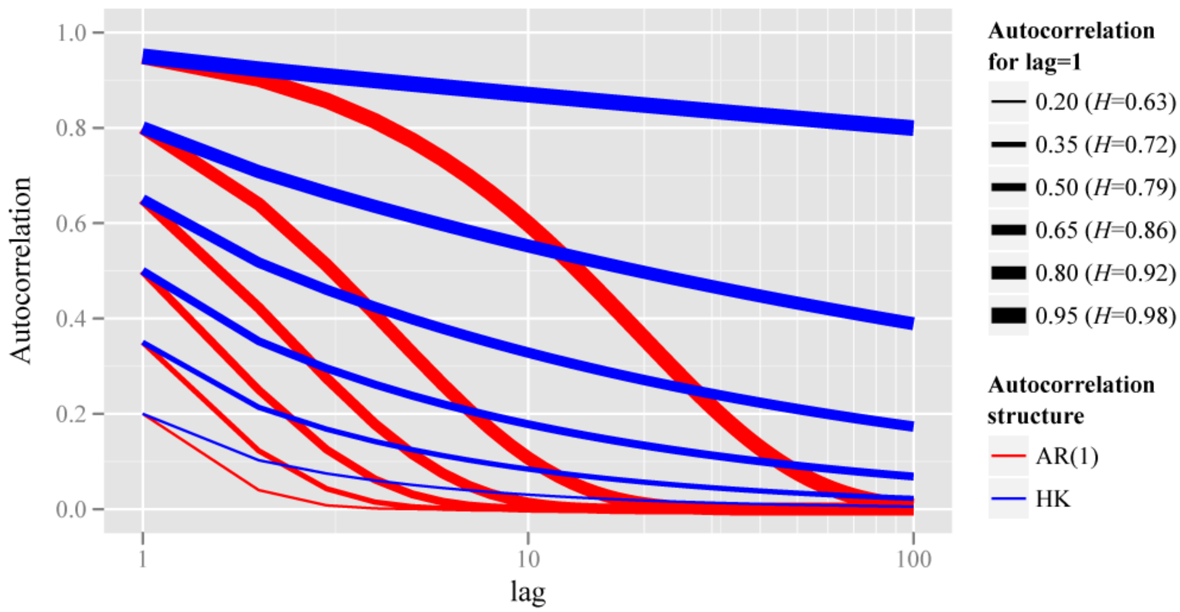


Figure 1. Theoretical autocorrelation functions for: (i) Markovian processes, AR(1), with exponential decay of autocorrelation with lag (Eq. (1)) and (ii) processes with HK behavior, described by the Hurst exponent *H*, with a power law relationship of autocorrelation with lag (Eq. (2)). The lag-1 autocorrelation, ρ, characterizes the strength of short-term persistence while the Hurst exponent, *H*, quantifies long-term dependences. Note that Eq. (2) implies that *H* and ρ are related as $H = 0.5\left[\log_2\left(\rho+1\right)+1\right]$.

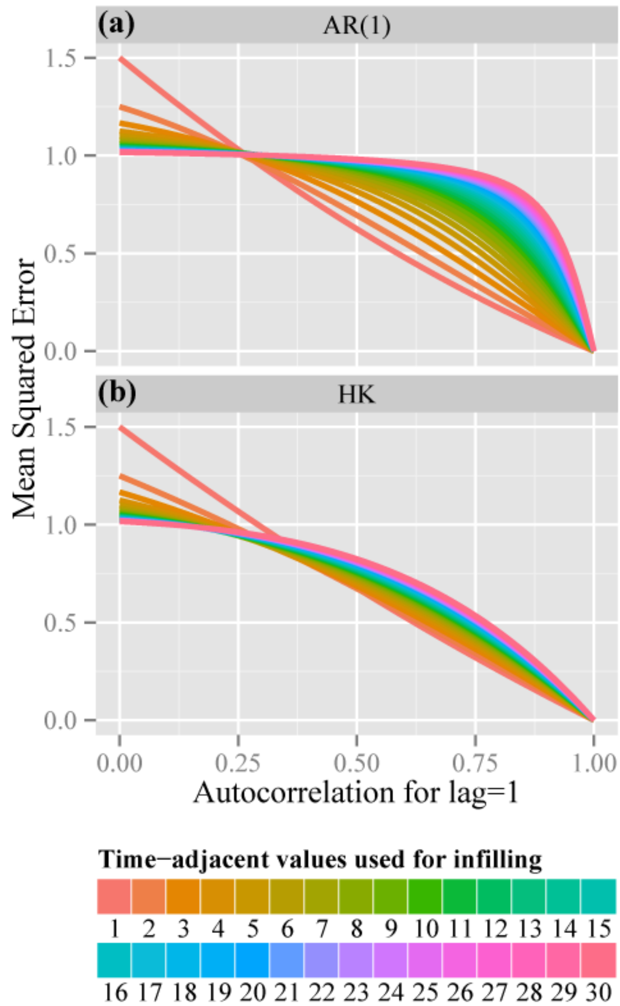


Figure 2. Illustration of the rationale underlying the Optimal Local Average (OLA) methodology based on Eq. (5) for processes with (a) exponential, and with (b) power-law autocorrelation structure. The Mean Squared Error of an estimated missing value, based on local averages with different range (i.e., different number of neighboring values), for hypothetical time series with different lag-1 autocorrelation and standard deviation equal to 1, is depicted (Eq. (5)). 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).

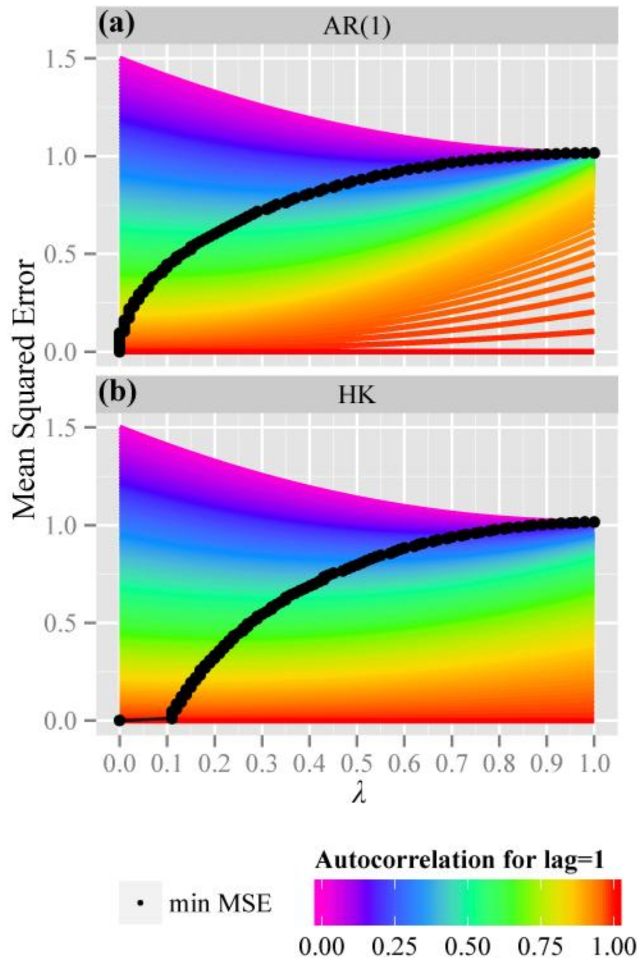


Figure 3. Surface plots of the Mean Squared Error (MSE) estimated according to the Weighted Sum of local and total Average (WSA) methodology (based on Eq. (8), and an hypothetical time series length of 2×30 and standard deviation equal to 1), for different values of parameter λ, for processes with (a) 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 values of parameter λ, which indicates the overall contribution of the global average, decreases.

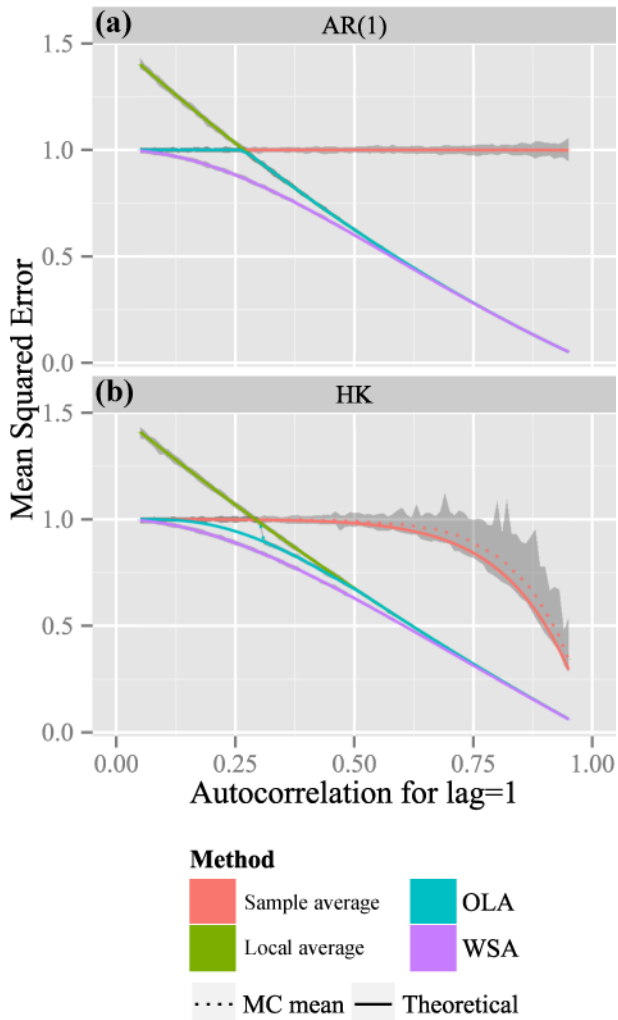
541 
(a)
AR(1)
(b)
HK
Mean Squared Error
0.0
0.5
1.0
1.5
0.00
0.25
0.50
0.75
1.00
Autocorrelation for lag=1
Method
Sample average
Local average
OLA
WSA
MC mean
Theoretical

Figure 4. 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; Optimal Local Average methodology, OLA; Weighted Sum of local and total Average approach (WSA). Results correspond to processes with (a) exponential, and (b) power-law (b) autocorrelation structure for different values of lag-1 autocorrelation. The solid lines depict the theoretical values of MSE (see Eq. (5) and Eq. (8)) while the dashed lines and uncertainty bounds correspond to the ensemble of the Monte-Carlo simulations, filling artificial data gaps. For the entire range of lag-1 autocorrelations, the WSA approach significantly outperforms other infilling methods, providing the smallest MSE.

(a)
AR(1)
Annual Precipitation [mm]
800
600
400
(b)
AR(1)
WSA (8 %)
OLA (2 %)
Local average (2 %)
Sample average (~0 %)
(c)
HK
Annual Temperature [°C]
10
9
8
7
1890
1920
1950
1980
2010
Year
(d)
HK
WSA (9 %)
OLA (2 %)
Local average (1 %)
Sample average (~0 %)
0%
2%
4%
6%
8%
1-nRMSE
(e)
HK
Minimum water depth [m]
6
4
2
800
850
900
950
1000
Year
(f)
HK
WSA (25 %)
OLA (22 %)
Local average (22 %)
Sample average (~0 %)
0%
5%
10%
15%
20%
25%
1-nRMSE

Figure 5. Real-world examples of time series with Markovian behavior (AR(1); annual precipitation, panel a) and with HK dynamics (annual temperature, panel c, annual minimum water depth, panel e). Original data are depicted in white circles, while the infilled time series are depicted in continuous colored lines. Each record was removed and infilled with the four examined approaches, calculating each time the new sample statistics. Bar-plots (panels b, d, f) illustrate the efficiency (defined as $1-\mathrm{nRMSE}$ where nRMSE is the normalized Root Mean Squared Error) of each gap-filling approach (i.e., sample average, local average, Optimal Local Average , OLA, and Weighted Sum of local and total Average, WSA). Since the sample average is re-calculated each time a value is removed, the efficiency of the sample average approach is not always equal to 0 (as expected theoretically, i.e., the nRMSE once the sample average should be 100 %).