Variability and Confidence Intervals for the Mean of Climate Data with Short- and Long-Range Dependence

MATTHEW C. BOWERS AND WEN-WEN TUNG

Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, Indiana

(Manuscript received 13 February 2017, in final form 17 March 2018)

ABSTRACT

This paper presents an adaptive procedure for estimating the variability and determining error bars as confidence intervals for climate mean states by accounting for both short- and long-range dependence. While the prevailing methods for quantifying the variability of climate means account for short-range dependence, they ignore long memory, which is demonstrated to lead to underestimated variability and hence artificially narrow confidence intervals. To capture both short- and long-range correlation structures, climate data are modeled as fractionally integrated autoregressive moving-average processes. The preferred model can be selected adaptively via an information criterion and a diagnostic visualization, and the estimated variability of the climate mean state can be computed directly from the chosen model. The procedure was demonstrated by determining error bars for four 30-yr means of surface temperatures observed at Potsdam, Germany, from 1896 to 2015. These error bars are roughly twice the width as those obtained using prevailing methods, which disregard long memory, leading to a substantive reinterpretation of differences among mean states of this particular dataset. Despite their increased width, the new error bars still suggest that a significant increase occurred in the mean temperature state of Potsdam from the 1896–1925 period to the most recent period, 1986–2015. The new wider error bars, therefore, communicate greater uncertainty in the mean state yet present even stronger evidence of a significant temperature increase. These results corroborate a need for more meticulous consideration of the correlation structures of climate data—especially of their long-memory properties—in assessing the variability and determining confidence intervals for their mean states.

1. Introduction

A time average is perhaps the most fundamental way to characterize climate, but the average or mean of climate data during some time intervals has limited utility without information about its variability. Useful constructs like error bars or confidence intervals—which can facilitate comparison with other periods, other locations, other scenarios, or between model and observation—are based on estimates of mean state variability. The challenge in characterizing variability is that the underlying correlation structure, or serial dependence, of the climate data must be properly considered, lest the resulting variability estimate be biased. Recognition of this problem in climate science dates back to the insightful work of [Leith \(1973\)](#), [Jones \(1975\)](#), and [Madden and Sadeh \(1975\)](#). Until very recently, approaches to account for serial dependence in estimates

of climate mean state variability have focused exclusively on short-range dependence, or short memory, which implies an exponential decay in a process's autocorrelation function (e.g., [Madden 1976](#); [Katz 1982](#); [Zwiers and von Storch 1995](#)).

However, another sort of temporal dependence structure called long-range dependence or long memory has since been detected in numerous physical state variables; for example, rainfall ([Lovejoy and Mandelbrot 1985](#); [Kantelhardt et al. 2006](#)), streamflow ([Hurst 1951](#); [Pelletier and Turcotte 1997](#); [Pandey et al. 1998](#)), tropical deep convection ([Tung et al. 2004](#)), general circulation ([Tsonis et al. 1999](#); [Vyushin and Kushner 2009](#)), surface temperature ([Lovejoy and Schertzer 1986](#); [Koscielny-Bunde et al. 1998](#); [Pelletier 1998](#); [Huybers and Curry 2006](#); [Vyushin and Kushner 2009](#); [Franzke 2010, 2012](#); [Yuan et al. 2015](#)), and even climate proxies like ice cores ([Schmitt et al. 1995](#); [Ashkenazy et al. 2003](#); [Thomas et al. 2009](#)), ocean cores ([Shackleton and Imbrie 1990](#)), and tree rings ([Bowers et al. 2013](#)). The long memory implies a slow power-law decay in a process's autocorrelation

Corresponding author: Matthew C. Bowers, mcbowers@purdue.edu

function, in contrast to the fast exponential decay of short-memory processes.

Short- and long-range dependence each cause a distinct effect on the variability of time averages, and both must be considered to obtain reasonable estimates of that variability (section 2). Thus, a procedure is needed that can incorporate both effects in estimates of mean state variability. To meet this need, we propose an adaptive and computationally feasible procedure for estimating the variance of time averages of climate data with short- and long-range dependence (section 3). The procedure is based on modeling the correlation structures in climate data with a parametric stochastic process, adaptively selecting among competing models, and estimating parameters using maximum likelihood estimation. The variance or standard error of mean states on a given time scale can then be computed analytically from the fitted model and used to construct confidence intervals. We illustrate the procedure by estimating variability and constructing confidence intervals for 30-yr time averages of the surface temperature at Potsdam, Germany (section 4). We provide evidence that interannual variability of the seasonal cycle is a source of long memory in the Potsdam temperature data (section 5). Discussion and comparison with related work are in section 6, and concluding remarks are given in section 7.

2. The effects of short and long memory on the variability of climate mean states

In this section, we provide evidence to support the claim that both short and long memory should be considered in the estimation of climate mean state variability. To do so, we first clarify the salient properties of short and long memories in terms of a process's temporal autocorrelation function and spectral density in the frequency domain. We then introduce the class of fractional autoregressive integrated moving-average (FARIMA) time series models, which can exhibit both short and long memories. Finally, we introduce the formula for the variance of time averages under short and long memories and use empirical simulation to establish intuition for the effects of the two distinct serial dependence structures.

a. Definition of short- and long-range dependence

Characterizing the variability of climate mean states is based on the knowledge of the process's correlation structure. For a climate process X_t with time index $t = 0, \dots, N$, mean μ , and variance σ^2 , the correlation structure is naturally expressed by the autocorrelation function

$$R(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2}, \quad (1)$$

where E is the expectation operator, and k is a time lag. The autocorrelation function $R(k)$ expresses the correlation between values of the process X_t that are separated by k time units. Previous studies of climate mean state variability have focused predominantly on short-memory autoregressive processes wherein each new value of X_t is modeled as a linear combination of a finite number of past values and noise (Leith 1973; Katz 1982; Zwiers and von Storch 1995). Such a stationary short-memory process exhibits exponential decay in its autocorrelation function $R(k) \sim \exp(-ak)$ for some positive constant a . There is extensive evidence of short memory in numerous atmospheric state variables, such as sea level pressure (Klein 1951; Madden 1976), surface temperature (Madden and Shea 1978), and geopotential height (Gutzler and Mo 1983).

More recently, evidence of long memory in various meteorological variables has also begun to accumulate; for example, precipitation (Lovejoy and Mandelbrot 1985; Kantelhardt et al. 2006), tropical deep convection (Tung et al. 2004), general circulation (Tsonis et al. 1999; Vyushin and Kushner 2009), and especially surface temperature (Koscielny-Bunde et al. 1998; Weber and Talkner 2001; Caballero et al. 2002; Eichner et al. 2003; Gil-Alana 2005; Huybers and Curry 2006; Vyushin and Kushner 2009; Franzke 2010, 2012; Yuan et al. 2015). A stationary long-memory process exhibits power-law decay in its autocorrelation function, that is, $R(k) \sim k^{2d-1}$ (where $0 < d < 1/2$ is the long-memory parameter), much slower than the exponential decay of a short-memory process. Furthermore, the summation over the autocorrelation function gives a measure of how quickly a perturbed process relaxes back toward its long-term mean state. Theoretically, for a short-memory process this summation is finite [$\sum_{k=-\infty}^{\infty} R(k) < \infty$], whereas for a long-memory process it diverges to infinity [$\sum_{k=-\infty}^{\infty} R(k) = \infty$] (e.g., Gao et al. 2006). This implies that a perturbed short-memory process returns toward its long-term mean state fairly rapidly, while a perturbed long-memory process may sustain the departure from its long-term mean for an extended time. This has the effect that the sequence of observed sample mean states from period to period of a short-memory process cluster more tightly around the long-term mean equilibrium state, while those of a long-memory process have greater spread or variability.

Another distinguishing characteristic of short- and long-memory processes manifests in the spectral density $f(\lambda) = (\sigma^2/2\pi) \sum_{k=-\infty}^{\infty} R(k) \exp(-ik\lambda)$ (where $i = \sqrt{-1}$ and $-\pi < \lambda < \pi$ is frequency). At low frequencies the spectral density theoretically takes the form

FIG. 1. (a)–(d) Raw Potsdam temperature data from the four 30-yr periods: 1896–1925, 1926–55, 1956–85, and 1986–2015. (e)–(h) Potsdam temperature anomalies after removal of the mean and seasonal cycle for the same periods.

$$f(\lambda) \sim c_f \lambda^{-2d}, \quad \lambda \rightarrow 0, \quad (2)$$

where c_f is a positive constant, $d=0$ implies short memory, and $0 < d < 1/2$ implies long memory with larger d indicating more intense long memory. This means that the spectrum of a long-memory process has a singularity at the origin, while that of a short-memory process converges to a finite value. This also means the long-memory parameter d conveniently distinguishes between short- and long-memory processes. Note that when the process has finite variance, the long-memory parameter d is related to the well-known Hurst parameter H (Hurst 1951) by $H = d + 1/2$.

In practice, correlation functions and power spectra are limited by sample length. To illustrate these properties in real meteorological data we analyze a long record of daily average surface temperatures. The measurements are taken from the meteorological station at the Potsdam Institute for Climate Impact Research located in Germany at 52.38°N, 13.07°E at an elevation of 100 m. The dataset, obtained from the German Weather Service Climate Data Center (2016, personal communication), consists of $N = 44\,924$ observations of daily average surface temperature spanning the 123 years from 1 January 1893 to 31 December 2015.

Figures 1a–d show the time series of the raw Potsdam temperature data over the four 30-yr periods: 1896–1925, 1926–55, 1956–85, and 1986–2015. The data exhibit a prevalent seasonal cycle that dominates the dependence structure as seen in their autocorrelation function (Fig. 2a). Despite the prevalence of seasonality, there is considerable temperature variability about the seasonal cycle. To scrutinize the more subtle dependence structure of this variability we must first separate it from the seasonal cycle.

This implies a simple model for atmospheric variability within a particular 30-yr time span similar to that in Katz (1982) in which observed data are considered as the superposition of a constant mean state, a seasonal cycle, potential nonstationary trends, and the remaining stochastic noise process. In addition to seasonality, if the data include significant nonstationary trends, they must also be removed prior to studying the dependence structure of the stationary stochastic process. For identification of trends in data with potential long memory, see, for example, Lennartz and Bunde (2009), Franzke (2010), Lennartz and Bunde (2011), and Ludescher et al. (2016). Adopting the methods of Franzke (2010), we do not find significant trends in the Potsdam data and therefore do not detrend the data (see section 4 for details). We remove the seasonal cycle by forward–backward

FIG. 2. (a) Autocorrelation function and (b) periodogram of the raw Potsdam temperature data in the most recent period 1986–2015. (c) Autocorrelation function and (d) periodogram of the Potsdam temperature anomalies after removal of the seasonal cycle. Negative autocorrelation values (6% of the 3650 correlations computed) are not shown in the log scale of (c). No detrending or tapering is used in constructing the raw periodograms in (b) and (d). The smoothed periodogram in (d) is obtained via a modified Daniell smoother in the frequency domain. Frequencies are given in cycles per year.

notch filtering the annual and semiannual cycles, obtaining the temperature anomaly series shown in Figs. 1e–h.

The autocorrelation function of these temperature anomalies, shown in Fig. 2c, indeed exhibits decay much slower than exponential—consistent with the power-law decay expected of long-memory processes. In addition, the periodograms of the raw and anomalous temperature data (Figs. 2b and 2d, respectively) appear to exhibit power-law scaling at low frequencies, also consistent with long memory. Despite these indications of long memory, the short time-scale segment of Fig. 2c is mostly linear with a hint of concavity, possibly due to short-memory effects. Indeed, such temperature data are known to

exhibit characteristics of short memory (e.g., Madden 1976; Katz 1982). With the indications of both short and long memories in the daily average surface temperature data, in the following subsection we review a class of parametric models that can simultaneously capture both dependence structures.

b. Fractional ARIMA models

We now introduce the class of FARIMA time series models, which can exhibit both short and long memories. This class is an extension of the classic short-memory autoregressive moving-average (ARMA) models introduced by Box and Jenkins (1970). The ARMA class has

been used in various meteorological and climatological applications including prediction of drought indices (Katz and Skaggs 1981), modeling quasi periodicity in zonal circulation (Jolliffe 1983), modeling precipitation occurrence and magnitude (Chang et al. 1984), stochastic parameterization of atmospheric convective heating (Lin and Neelin 2002), and wind forecasting (Erdem and Shi 2011; Ailliot and Monbet 2012). The ARMA class models the values of a temporal process X_t sampled uniformly in time ($t = 1, \dots, N$) as a linear combination of past values plus the linear combination of current and previous stochastic noise innovations, that is,

$$X_t = \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}, \quad t = 1, \dots, N, \quad (3)$$

where $\phi_1, \dots, \phi_p$ are parameters called the autoregressive coefficients, $\theta_1, \dots, \theta_q$ are parameters called the moving-average coefficients, ε_t ($t = 1, \dots, n$) is an uncorrelated Gaussian noise process with mean zero and variance σ_ε^2 , and X_t is called an ARMA(p, q) process.

Box and Jenkins (1970) also extended the ARMA class by introducing integrated autoregressive moving-average (ARIMA) processes. For instance, the cumulative sum Y_t of an ARMA(p, q) process X_t can be defined with $X_t = Y_t - Y_{t-1}$. To make the notation compact and allow for various differencing/integration orders, Box and Jenkins (1970) used a backshift operator B , which imparts a lag to a temporal process; that is, $B^k X_t = X_{t-k}$ for integer k . Then the cumulative sum Y_t of X_t can be written as $(1 - B)Y_t = X_t$, and more generally the order- m integration can be written as $(1 - B)^m Y_t = X_t$, where X_t is ARMA(p, q) and Y_t is ARIMA(p, m, q).

Granger and Joyeux (1980) and Hosking (1981) realized that by allowing the integer differencing order m of the ARIMA models to be fractional d , a new process could be obtained that generalizes both the stationary ARMA and nonstationary ARIMA processes. A FARIMA(p, d, q) process Y_t is defined recursively by $Y_0 = 0$ and

$$(1 - B)^d Y_t = X_t, \quad (4)$$

where X_t is an ARMA(p, q) process as defined by Eq. (3), and the differencing order d is a real number, identical to the long-memory parameter d discussed in section 2a. Thus, the parameter d controls the long-memory intensity with $0 < d < 1/2$ corresponding to long memory and $d = 0$ corresponding to short memory. The short-memory dependence structure is controlled by the q moving-average coefficients $\theta_1, \theta_2, \dots, \theta_q$

and the p autoregressive coefficients $\phi_1, \phi_2, \dots, \phi_p$ [Eq. (3)].

The class of FARIMA(p, d, q) models is particularly useful because it encompasses processes with pure short memory, pure long memory, and simultaneous short and long memory. For instance, when $d = 0$, the FARIMA(p, d, q) reduces to the short-memory ARMA(p, q) model, and when $q = 0$, it further reduces to the autoregressive AR(p) model. Since they generalize ARMA models by including the possibility of long memory, there is no surprise that FARIMA models have been used in numerous meteorological and climatological applications; for example, characterizing properties of rainfall (Montanari et al. 1996), describing deep-ocean variability (von Storch et al. 2001), modeling surface temperatures (Caballero et al. 2002; Franzke 2010), and forecasting wind speed (Kavasseri and Seetharaman 2009).

c. Variance of time averages under short and long memories

We now support the claim that both short and long memories should be considered in the estimation of climate mean state variability. We first present a general formula for the variance of the sample mean of stationary processes in the FARIMA(p, d, q) class having short-, long-, or both short- and long-range dependence. We then clarify the interpretation of this formula empirically through a comparison of the Potsdam temperature anomalies and four simulated FARIMA processes, distinguishing the respective effects of short and long memories on the variance of time averages.

Consider a stationary FARIMA(p, d, q) process Y_t ($t = 1, 2, \dots, N$) with true underlying mean μ . Even under serial dependence, the sample mean of size n , $\bar{Y}_n = \sum_{t=1}^n Y_t/n$, $n \leq N$, is an unbiased estimator of the true underlying process mean μ ; that is, the expectation of $\bar{Y}_n$ is equal to μ . Of course, for a given sample, $\bar{Y}_n$ will not be exactly equal to μ ; in fact $\bar{Y}_n$ follows a Gaussian distribution with mean μ and variance $\text{Var}[\bar{Y}_n]$. For large n , the general formula for the variance of $\bar{Y}_n$ is (Beran et al. 2013, chapter 1)

$$\text{Var}[\bar{Y}_n] \sim \nu(d)f(n^{-1})n^{-1} \sim \nu(d)c_f n^{2d-1}, \quad n \rightarrow \infty, \quad (5)$$

where

$$\nu(d) = \begin{cases} \frac{2\Gamma(1-2d)\sin(\pi d)}{d(2d+1)}, & d \neq 0, \\ 2\pi, & d = 0. \end{cases}$$

As $n \rightarrow \infty$, $f(n^{-1})$ corresponds with the spectral density at low frequencies, which takes the form [cf. Eq. (2)]

$$f(\lambda) \sim \frac{\sigma_\varepsilon^2}{2\pi} \left| \frac{1 + \sum_{j=1}^q \theta_j}{1 - \sum_{i=1}^p \phi_i} \right|^2 |\lambda|^{-2d} \sim c_f |\lambda|^{-2d}, \quad \lambda \rightarrow 0, \quad (6)$$

so that c_f is the prefactor of the spectral density at low frequencies, containing the information about the short-memory autoregressive and moving-average components of the process. We show in [appendix A](#) that this formulation generalizes the large sample approximation for computing variances of time averages of short-memory processes used in previous climatological literature.

To clarify the interpretation of Eq. (5), we compute the empirical variance–time relations (e.g., [Tung et al. 2004](#)) between various averaging sizes n and $\text{Var}[\bar{Y}_n]$ for the Potsdam temperature anomalies and four different simulated FARIMA series. The empirical variance–time relation is obtained by partitioning a length N series into nonoverlapping blocks of length $n \ll N$ samples, computing the sample mean inside each block, and computing the sample variance among these block means. While the Potsdam temperature data have $N = 44\,924$ samples, the four simulated FARIMA time series have a length of $N = 10^6$ each and have the following dependence structures:

- 1) uncorrelated white noise FARIMA(0, 0, 0);
- 2) short-memory FARIMA(3, 0, 0), AR(3), with $\phi_1 = 0.76$, $\phi_2 = -0.16$, and $\phi_3 = 0.06$;
- 3) long-memory FARIMA(0, d , 0) with $d = 0.16$; and
- 4) simultaneous short- and long-memory FARIMA(3, d , 0) with $d = 0.16$, $\phi_1 = 0.76$, $\phi_2 = -0.16$, and $\phi_3 = 0.06$.

The four simulated series are then normalized to have unit variance.

[Figure 3](#) shows the estimated variance–time curves for the Potsdam temperature anomalies and the four simulated FARIMA series. Since the series are normalized, the variance–time curves all originate with unit variance at $n = 1$; however, they assume different shapes depending on their short- and long-range dependence structures. As expected from the classical central limit theorem, the variance–time curve of the white noise series decays as n^{-1} for all n , appearing as a line with slope of -1 in the log–log scale. For short memory ($d = 0$), Eq. (5) indicates a variance–time curve of $2\pi c_f n^{-1}$ as $n \rightarrow \infty$. Indeed, the empirical variance–time curve for the AR(3) decays asymptotically as n^{-1} . But the short memory causes slower decay at small n , which effectively scales the asymptotic variances by the factor

FIG. 3. Empirical variance–time relations between sample variance of block means and block size for the Potsdam temperature anomalies, independent white noise, short-memory AR(3), long-memory FARIMA(0, d , 0), and both short- and long-memory FARIMA(3, d , 0).

of $2\pi c_f$, manifesting as a vertical shift above the white-noise curve in the log–log scale.

According to Eq. (5), a long-memory process has an asymptotic variance–time relation of $\nu(d)c_f n^{2d-1}$. This tendency to decay slower than n^{-1} is a fingerprint of long memory known as the Joseph effect or Hurst phenomenon ([Mandelbrot and Wallis 1968](#)). Indeed, the pure long-memory FARIMA(0, d , 0) variance–time curve decays as $n^{2d-1} = n^{2(0.16)-1} = n^{-0.68}$, appearing as a line with a slope of -0.68 in the log–log scale. The variance–time relation of the short- and long-memory FARIMA(3, d , 0) exhibits both dependence effects with a slow asymptotic decay rate of $n^{2d-1} = n^{-0.68}$ and scaling by the factor $\nu(d)c_f$, which shifts it above the FARIMA(0, d , 0) curve. The variance–time curve of the Potsdam temperature anomalies coincides well with that of the FARIMA(3, d , 0), exhibiting both the slow asymptotic decay of long memory and the vertical shift of short memory. Neither a pure short-memory nor a pure long-memory model can mimic the variance–time relation of the Potsdam temperature anomalies, but a model combining both dependence structures can do so quite well.

From Eq. (5) and these empirical variance–time relationships, we can see that at large averaging sample sizes, short-range dependence scales the variance by a constant factor (vertical shift in log–log scale), while long-range dependence actually changes the rate of decay (less steep slope in log–log scale). This means that

if one uses a pure short-memory model to represent the variance–time relation of climate data that actually have long memory, variances of mean states on large time scales will be underestimated. Thus, in order to reliably quantify the variability of climate mean states, one must detect and characterize *both* the short- and long-range correlation structures of the climate data.

3. Characterizing the variability of climate mean states under short and long memories

In this section, we describe an adaptive and computationally feasible procedure for estimating the variance and constructing confidence intervals for time averages of climate data with short- and long-range dependence. The procedure is based on modeling climate data as a FARIMA(p, d, q) process, estimating parameters using approximate maximum likelihood estimation, and adaptively selecting the model order using an information criterion. The variance or standard error of mean states on a given time scale can then be computed directly from the fitted model using Eq. (5).

a. Preliminary steps

Before applying the forthcoming variance estimation procedure, several preliminary steps should be taken to ensure its application to a particular dataset is appropriate. First, any prevalent seasonality should be removed—for example, the annual cycle and, if substantial, its first few harmonics. The forthcoming variance estimation procedure is formulated for stationary linear Gaussian processes, so after removing the seasonal cycle, data should be at least roughly consistent with these conditions. Nevertheless, as we discuss below, the procedure is expected to be robust against non-Gaussian data.

Exploratory and heuristic methods can be applied to assess any evidence for short and/or long memories. Short memory can be assessed through plots of the autocorrelation function or the variance–time relation. The intensity of any possible long memory can be assessed using any of the numerous heuristic methods, including the rescaled range method (Hurst 1951), the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) statistic (Kwiatkowski et al. 1992), the rescaled variance (V/S) method (Giraitis et al. 2003), the detrended fluctuation analysis (DFA; Peng et al. 1994), Haar wavelets (Lovejoy and Schertzer 2012), and the adaptive fractal analysis (Gao et al. 2011). If heuristic methods are indicative of long memory, then a more complete characterization of the process's dependence structure may be obtained through the more rigorous technique of maximum likelihood estimation.

b. Model estimation

FARIMA time series models can be fit to data using maximum likelihood estimation, which is a method for estimating the parameter values of a statistical model's given data. The method works by searching for the parameter values that maximize the likelihood function, a measure of the degree to which the data support particular parameter values (e.g., Lindgren 1976).

Fitting FARIMA models using maximum likelihood estimation has two major advantages in this context: it allows for the simultaneous estimation of both short- and long-memory structures, and it allows the subsequent use of information-based criteria to select among competing models. However, exact maximum likelihood estimation is computationally infeasible because of numerous inversions of the $n \times n$ covariance matrix. Fortunately, numerous approximate likelihood methods have been developed, including both Frequentist (e.g., Beran et al. 2013, chapter 5) and Bayesian (e.g., Graves et al. 2015) approaches. Here, we employ the elegant and computationally efficient spectral domain approximation to the Gaussian likelihood proposed by Whittle (1953). The Whittle estimator is asymptotically equivalent in distribution to the exact maximum likelihood estimator for Gaussian data (Fox and Taqqu 1986) and yields asymptotically consistent and normally distributed parameter estimates for non-Gaussian data (Giraitis and Surgailis 1990). Simulation studies have confirmed that the Whittle estimator is indeed robust against non-Gaussian data and outperforms other popular methods under both short- and long-range dependence (Taqqu and Teverovsky 1998; Franzke et al. 2012).

The Whittle method is based on the periodogram of the process Y_t ($t = 0, 1, \dots, N$),

$$I(\lambda) = \frac{1}{2\pi N} \left[\sum_{t=0}^N Y_t \exp(it\lambda) \right]^2, \quad (7)$$

where $-\pi < \lambda < \pi$ is frequency, and $i = \sqrt{-1}$. The method relies on minimizing the function

$$Q(\boldsymbol{\eta}) = \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda; \boldsymbol{\eta})} d\lambda + \int_{-\pi}^{\pi} \log f(\lambda; \boldsymbol{\eta}) d\lambda, \quad (8)$$

where $\boldsymbol{\eta}$ is the vector of unknown parameters, which contains the long-memory parameter d as well as any autoregressive and moving-average coefficients [see Eq. (3)], and $f(\lambda; \boldsymbol{\eta})$ is the spectral density function given the parameters in $\boldsymbol{\eta}$. The Whittle estimator is defined as the value of $\boldsymbol{\eta}$ that minimizes $Q(\boldsymbol{\eta})$; for computation, the integrals in Eq. (8) are replaced by the corresponding sums over Fourier frequencies.

c. Model selection

Maximum likelihood–based methods like the Whittle estimation require specification of the precise parametric form of the model; that is, the value of p and q and whether or not $d = 0$, which can invite bias under model misspecification (e.g., Fox and Taqqu 1986). Thus, it is advisable to use a model selection procedure to identify an appropriate order of the FARIMA(p, d, q) model.

What constitutes an appropriate model depends on the aim of the investigation. Since our ultimate purpose is to quantify the variability in time averages of climate data, we are primarily concerned with satisfactorily representing the process's dependence structure. Thus, we use Occam's razor or parameter parsimony as a guiding principle in model building. This proposition suggests that among adequate models, the one with the fewest parameters is preferable.

To identify the preferred model, that is, the preferred values of p and q , we use an information-based selection procedure. The idea is to begin with a pool of candidate models, such as FARIMA(p, d, q) models with $p \leq p_{\max}$, $q \leq q_{\max}$, and $1/2 < d < 1/2$; since we seek a parsimonious model, we find $p_{\max} = q_{\max} = 8$ adequate. We make an information-based comparison among these candidate models to identify one that is best in terms of minimizing information loss (e.g., Akaike 1974). We follow Beran et al. (1998), who found that the Bayesian information criterion (BIC) of Schwarz et al. (1978) performs well in the parsimonious selection of FARIMA models. The BIC may be interpreted as a measure of the information lost by the use of a fitted model rather than the data themselves (Burnham and Anderson 2002). The BIC is defined as

$$\text{BIC} = -2 \log(\tilde{L}) + \log(n) \times r, \quad (9)$$

where $\tilde{L}$ is the maximized value of the likelihood function, that is, the likelihood of the Whittle method parameter estimates, n is the number of samples used in the estimation, and r is the number of parameters being estimated. The $\log(n) \times r$ term encourages parsimony by acting as a penalty for adding additional terms to the model. We first narrow the candidate pool to those models for which Whittle's parameter estimation procedure converges, and we identify the model minimizing BIC as the most preferred model in the candidate pool.

d. Confidence intervals for climate mean states

In this section, we describe how the variability estimates presented in this paper may be used to construct a

confidence interval for the mean of climate data. A confidence interval quantifies our knowledge about the true mean state by bracketing a set of plausible values, based on a sample sequence of data. Confidence intervals can therefore serve as a convenient error bar for estimated climate mean states.

Suppose we are interested in the mean of a sequence of climate data Y_t during some time span of length n . It is assumed that the sample mean $\bar{Y}_n$ has been estimated and subtracted from the data, that the seasonal cycle has been removed, and that the FARIMA(p, d, q) model has been selected and fit to the anomalous data via the procedures in sections 3b and 3c. A confidence interval for the mean μ of Y_t is essentially the sample mean plus or minus the appropriate quantile from a standard normal distribution times the square root of the variance of the sample mean. We estimate the variance of the sample mean with Eq. (5), replacing d and f with their estimates, $\hat{d}$ and $\hat{f}$; the latter is the spectral density in Eq. (6) given the estimated FARIMA(p, d, q) parameter values.

Then, for large n , a two-sided $(1 - \alpha) \times 100\%$ confidence interval for μ is

$$\bar{Y}_n \pm z_{1-\alpha/2} \sqrt{\nu(\hat{d})\hat{f}(n^{-1})n^{-1}}, \quad (10)$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution (Beran et al. 2013, chapter 5). The one-sided upper $(1 - \alpha) \times 100\%$ confidence bound is

$$\bar{Y}_n + z_{1-\alpha} \sqrt{\nu(\hat{d})\hat{f}(n^{-1})n^{-1}}, \quad (11)$$

and the lower $(1 - \alpha) \times 100\%$ confidence bound is

$$\bar{Y}_n - z_{1-\alpha} \sqrt{\nu(\hat{d})\hat{f}(n^{-1})n^{-1}}. \quad (12)$$

In practice, since we are estimating $\hat{f}$, the normal approximation may not be very accurate at small sample sizes because of uncertainty in the estimated parameters. This is analogous to the more familiar situation for independent data in which a t distribution is used when the variance is unknown and must be estimated. Beran et al. (2013) use a Monte Carlo simulation to estimate the additional variability induced by parameter uncertainty. For a 95% confidence interval, the $z_{1-\alpha/2}$ factor should be inflated by about 5% when parameters are estimated from a sample size of 1000. Again, analogous to the situation with independent data, for very large sample sizes, such as the $>10^4$ used in section 4, this inflation is essentially negligible.

FIG. 4. Schematic diagram illustrating the procedure for obtaining appropriate error bars for the mean of climate time series data. Enumeration corresponds to the tasks numbered in the text.

4. Analysis procedure demonstrated with the Potsdam surface temperature record

In this section, we demonstrate a procedure for the determination of error bars on the mean of climate data that may have both short- and long-range dependence. We illustrate the procedure by determining error bars for mean states of the 123-yr record of observed daily average surface temperatures at Potsdam, Germany. Since an average over at least 10 years of daily observations for most common state variables (30 years for precipitation) is a classical climate definition by the World Meteorological Organization ([World Meteorological Organization 2011](#)), we focus on mean states of the four 30-yr periods 1896–1925, 1926–55, 1956–85, and 1986–2015. Specifically, we estimate the time mean for each period along with its variance or standard error and construct a 95% confidence interval for the error bars of each period.

The procedure for obtaining appropriate error bars for the mean of climate time series data is represented schematically in [Fig. 4](#). The salient tasks are

- 1) remove the seasonal cycle and nonstationary trends;
- 2) assess the potential for long memory with heuristic methods such as DFA or the variance–time relation;
- 3) fit a set of candidate short-memory ARMA models and select the best one according to the BIC;
- 4) fit a set of candidate long-memory FARIMA models and select the best one according to the BIC;

- 5) choose between the best ARMA and best FARIMA models using diagnostic visualization or summative tests; and
- 6) determine the error bars for the time mean based on the chosen model.

We illustrate these tasks by independently applying them to each 30-yr segment of data, beginning with the removal of the seasonality by forward–backward notch filtering the annual and semiannual cycles.

The next task is to remove any nonstationary trends that exist in the data. The issue of trend removal should be treated with care, as even a stationary stochastic process can appear to have trend, especially if it has long memory ([Bunde et al. 2014](#); [Ludescher et al. 2016](#)). Removing arbitrary trends from data that are actually stationary is no safer than treating nonstationary data as stationary; therefore, we detrend data only given compelling evidence that a trend exists.

To detect possible trends, we adopt the approach of [Franzke \(2010\)](#). For each segment, we first remove a linear trend and then choose a best FARIMA model by the methods in [section 3c](#). We then simulate an ensemble of 1000 time series from the fitted model and estimate the magnitude of linear trend in each synthetic time series. Finally, we compare the magnitude of the linear trend in the observed data with the distribution of trend magnitudes from the simulated data and compute p values. None of these p values are significant at $\alpha = 0.05$, so we conclude there is no need to detrend any

FIG. 5. Variance–time plot of the Potsdam temperature anomalies within each of the 30-yr periods on record. The regression line is fitted to scales $2^5 \leq m \leq 2^{10}$ (32–1024 days). The estimates $\hat{d}_{VT}$ are based on the slope, which corresponds to $2d - 1$, and the standard deviations of the estimates are given in parentheses.

period of the Potsdam data. We return to the issue of arbitrary trends in the data and its relationship to estimation of long memory in [section 5](#).

To assess the potential for long memory in the data, we check the variance–time relation in [Fig. 5](#) and the first-order DFA ([Peng et al. 1994](#)) in [Fig. 6](#). Since $\text{Var}[\bar{Y}_n]$ is expected to decay as n^{2d-1} ([section 3c](#)), the asymptotic slope of the variance–time relation admits a heuristic estimate of long-memory intensity $\hat{d}_{VT}$. The resulting estimates of the long-memory parameter, $\hat{d}_{VT}$ and $\hat{d}_{DFA}$, are given in [Table 1](#). Both DFA and the variance–time relations are consistent with long memory in the four time periods, with the exception of DFA for 1896–1925, which is more consistent with pure short memory. While they

can be useful for exploring the potential for long memory in climate data, heuristic methods like DFA are known to be biased in the presence of short-range dependence ([Taqqu and Teverovsky 1998](#)). For rigorous parametric modeling, we use instead a maximum likelihood–based approach, which can simultaneously characterize both short and long memories.

To select the best ARMA and the best FARIMA models for the data, we first consider a pool of candidate $\text{ARMA}(p, q)$ models and a pool of candidate $\text{FARIMA}(p, d, q)$ models all having $p \leq p_{\max}$ and $q \leq q_{\max}$. With parsimony as the guiding principle of selection, we find $p_{\max} = q_{\max} = 8$ to yield a sufficiently wide pool of candidate models. For each 30-yr period,

FIG. 6. First-order DFA (DFA-1) of the Potsdam temperature anomalies within each of the 30-yr periods on record. The regression line is fit to scales $w \geq 2^5$ (32 days) yielding the estimates $\hat{d}_{DFA}$ and their standard deviations (in parentheses).

we fit 81 candidate ARMA models and 81 candidate FARIMA models using the Whittle method (see [section 3b](#)), and we select the best model from each pool according to the BIC (see [section 3c](#)). The parameter estimates for the selected ARMA and FARIMA models for each 30-yr period are shown in [Tables 2 and 3](#), respectively.

[Figure 7](#) shows the periodogram for each 30-yr period of the Potsdam temperature anomalies along with the spectra of the selected FARIMA and ARMA models. In each period, the BIC selects low-order models with totals of at most four moving-average and autoregressive parameters. Differences between the ARMA and

FARIMA spectra are most evident at low frequencies where the FARIMA spectra have greater amplitude associated with the presence of long memory.

The next task is to choose between the best candidate ARMA and FARIMA models. For this task, we adopt a visual diagnostic plot that provides a comprehensive view of model quality across various scales of atmospheric variability. Several summative approaches are also available, including the difference of BIC (ΔBIC), goodness-of-fit testing, the likelihood-ratio test, and the simulation-based selection strategy of [Rust \(2007\)](#) (see [appendix B](#)). We use the visual diagnostic at this step because the summative approaches may favor a model

TABLE 1. Heuristic estimates of the long-memory parameter of the average daily surface temperature record in each 30-yr period where $\hat{d}_{VT}$ and $\hat{d}_{DFA}$ are obtained from the variance–time relation and the detrended fluctuation analysis, respectively. Standard deviations of the estimates based on linear regression are given in parentheses.

Period	$\hat{d}_{VT}$	$\hat{d}_{DFA}$
1896–1925	0.06 (0.02)	−0.02 (0.03)
1926–55	0.10 (0.02)	0.15 (0.01)
1956–85	0.08 (0.02)	0.06 (0.02)
1986–2015	0.15 (0.02)	0.07 (0.03)

for its overall better fit across time scales without generally revealing its scale- or range-dependent strengths or deficiencies.

To create a diagnostic visualization capable of revealing model quality across various regimes of atmospheric variability, we employ the average–transform–smooth (ATS) method described by Cleveland et al. (1993). The procedure illuminates the residuals between the raw periodogram $I(\lambda)$ of a time series and the spectrum $f(\lambda)$ of a fitted model. First, the periodogram and fitted spectrum are averaged within nonoverlapping blocks of 4 points, each yielding $\bar{I}(\lambda)$ and $\bar{f}(\lambda)$, and the average values are assigned to the average frequency of the block. Cleveland et al. (1993) found that a block size as small as 4 still provides excellent statistical properties. Next, a variance-stabilizing natural log transformation is applied to the quotient $\bar{I}(\lambda)/\bar{f}(\lambda)$ to obtain the residuals $\ln[\bar{I}(\lambda)] - \ln[\bar{f}(\lambda)]$. These residuals represent discrepancies between the empirical spectrum of the data and the spectrum of the fitted model.

Figure 8 shows the diagnostic spectral residual plots for both ARMA and FARIMA models over the four 30-yr periods of the Potsdam temperature data. The residuals are grouped into three frequency bands corresponding to distinct regimes of atmospheric variability: from 2 days to 2 weeks, from 2 weeks to 1 year, and from 1 to 30 years. The high-frequency band from 2 days to 2 weeks corresponds with mesoscale to synoptic variability; the middle band from 2 weeks to 1 year corresponds with the subseasonal to seasonal scale as defined in National Academies of Sciences, Engineering, and Medicine (2016); and the low-frequency band corresponds with interannual to multidecadal scales. The distributions of residuals in each frequency band are summarized by a boxplot; residual distributions not centered at zero imply bias in the model in that frequency band. The four panels indicate that both ARMA and FARIMA models perform reasonably well at time scales below 1 year. However, for the estimation of variance at long time scales, aptness at low frequencies is critical. Figure 8a shows that, in the 1896–1925 period,

TABLE 2. Parameter estimates for the ARMA model [see Eq. (3)] minimizing the BIC in each 30-yr period.

Period	P	q	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2
1896–1925	1	2	0.759	—	—	0.134	−0.084
1926–55	3	0	0.900	−0.193	0.077	—	—
1956–85	2	2	1.451	−0.505	—	−0.496	−0.217
1986–2015	2	2	1.509	−0.543	—	−0.558	−0.226

while the ARMA model performs well, the FARIMA model overestimates variability on time scales longer than 1 year, resulting in a negative residual distribution in that band. Figures 8b–d show that while the FARIMA performs well in all frequency bands, the ARMA model underestimates variability at low frequencies, resulting in positive bias in the residual distributions. These findings are consistent with the results of DFA shown in Fig. 6, which indicated long memory in the latter three periods. Based on these results, we conclude that the ARMA model is preferred for the 1896–1925 period, while FARIMA models are preferred for the other three periods.

Given the chosen models, we can determine appropriate error bars for the time mean of each period by using Eq. (10) to compute 95% confidence intervals. We can contrast the FARIMA-based confidence intervals with their status quo ARMA-based counterparts to understand if the difference in uncertainty characterization is meaningful. Table 4 and Fig. 9 show the time mean of each 30-yr period along with the 95% confidence intervals computed from both the selected ARMA and FARIMA models. Although the width of the error bars varies across the four periods, those from the FARIMA models, which account for long memory, are consistently wider than those from the ARMA models, which do not. In fact, the distinction is substantive; for example, a comparison of the mean temperature states for the periods of 1956–85 and 1986–2015 informed by ARMA-based error bars indicates a significant difference (the two confidence intervals do not overlap), whereas a comparison informed by FARIMA-based error bars indicates no significant difference (the two confidence intervals overlap). Somewhat paradoxically, while the wider error bars decrease the apparent significance of changes in mean temperature from period to period, they also communicate greater uncertainty, which means that the true difference in means could be much greater than previously thought. Bunde et al. (2014) and Ludescher et al. (2016) reported similar findings in which increased uncertainty due to long memory simultaneously makes west Antarctic warming trends less significant yet allows for much larger trends than previously thought.

TABLE 3. Parameter estimates for the FARIMA [see Eqs. (3) and (4)] model minimizing the BIC in each 30-yr period. SE denotes the standard deviation of $\hat{d}$, estimated based on the maximum likelihood.

Period	P	q	$\hat{d}$	$SE[\hat{d}]$	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2
1896–1925	1	2	0.095	0.030	0.688	—	—	0.110	−0.084
1926–55	3	0	0.143	0.029	0.753	−0.145	0.055	—	—
1956–85	3	0	0.119	0.031	0.836	−0.200	0.068	—	—
1986–2015	3	0	0.193	0.029	0.759	−0.173	0.066	—	—

Comparing the periods 1926–55 and 1986–2015 leads to the same contradiction between ARMA- and FARIMA-based inferences. Only the comparison of periods 1896–1925 and 1986–2015 leads to a unanimous conclusion of change in mean temperature from both ARMA- and FARIMA-based error bars. Such substantive

discrepancies between conclusions emphasize the necessity for a meticulous choice of the model on which confidence intervals are based.

With the emergence of climate change as a major public policy issue, better characterization of uncertainty has become increasingly critical. Improvements in

FIG. 7. Periodogram of the Potsdam temperature anomalies along with the spectra of the selected FARIMA and ARMA models for each 30-yr period on record. Estimates of the long-memory parameter d included in the FARIMA models ($\hat{d}$) are given along with their standard deviations in parentheses.

FIG. 8. ATS-based spectral diagnostic visualization for the models fitted to the four 30-yr periods of Potsdam temperature anomalies. The plots show the distribution of spectral residuals $\ln[\hat{I}(\lambda)] - \ln[\hat{f}(\lambda)]$ corresponding to both the selected ARMA and FARIMA models for each period.

uncertainty characterization have emerged for both observational datasets (e.g., [Santer et al. 1999](#)) and climate model simulations (e.g., [Majda and Gershgorin 2010](#)). The procedure presented here offers an improvement applicable to both observational and model data by allowing for a more faithful representation of the variability and dependence structure of the data on which error bars are based. While the FARIMA-based confidence intervals are wider than their ARMA-based counterparts and therefore communicate greater uncertainty in

the time mean, they provide even stronger evidence of the increase in mean temperature in the most recent 30-yr period.

5. Sources of long memory in the Potsdam temperature data

In this section, we explore the source of long memory in the Potsdam temperature data. In particular, we focus on the potential role that nonstationarity, seasonality,

TABLE 4. Mean daily surface temperature at Potsdam for each of the four 30-yr periods ($\bar{Y}_{30\text{yr}}$; °C) along with the variance of the mean ($\text{Var}[\bar{Y}_{30\text{yr}}]$), the half-width of the 95% confidence interval $\{\text{CI}; z_{0.975} \times \text{Var}[\bar{Y}_{30\text{yr}}]^{1/2}$; see Eq. (10)), and lower and upper bounds of the 95% CI estimated from the selected ARMA and FARIMA models.

Period	Model	$\bar{Y}_{30\text{yr}}$	$\text{Var}[\bar{Y}_{30\text{yr}}]$	95% CI		
				Half-width	Lower	Upper
1896–1925	ARMA(1, 2)	8.420	0.009	0.190	8.229	8.610
	FARIMA(1, d , 2)		0.030	0.339	8.080	8.759
1926–55	ARMA(3, 0)	8.763	0.011	0.203	8.559	8.966
	FARIMA(3, d , 0)		0.061	0.484	8.278	9.247
1956–85	ARMA(2, 2)	8.668	0.012	0.219	8.449	8.887
	FARIMA(3, d , 0)		0.044	0.410	8.259	9.078
1986–2015	ARMA(2, 2)	9.534	0.018	0.264	9.270	9.798
	FARIMA(3, d , 0)		0.131	0.710	8.824	10.244

FIG. 9. Mean states of the Potsdam temperature along with the 95% confidence intervals obtained assuming short memory (ARMA model) and accounting for long memory (FARIMA model) for each 30-yr period on record.

and arbitrary trends in the time series data could play in estimating the intensity of long memory.

To simultaneously identify the seasonal cycle and potential trends, we use seasonal decomposition of time series using Loess (STL; Cleveland et al. 1990), a nonlinear filtering procedure for decomposing a time series into trend, seasonal, and remainder components.

Figure 10 shows the STL decomposition of the Potsdam temperature data in the 30-yr period from 1986 to 2015. For visibility, the components are plotted on different vertical scales, and rectangles spanning the same temperature range are provided on the right side of the plots for comparison. STL recovers an unambiguously increasing trend as well as an annual cycle that is allowed to vary from period to period, consistent with the interannual variation found in the seasonal cycle of climate data. After removing the trend and seasonality, we are left with a remainder series of correlated noise, which is stationary in the mean.

We repeat the analysis procedure described in section 4 using the STL detrended and deseasoned data. After selecting a candidate FARIMA and ARMA model for each period, we consult the spectral residual diagnostic plots and find that FARIMA models are no longer superior to the ARMA models on time scales beyond 1 year. This implies that removal of the STL seasonal and trend components effectively removed long memory from the Potsdam data.

To further identify the main source of long memory, we add the STL trend components back into the stationary remainder components, effectively deseasoning but not detrending the data, and we repeat the analysis of section 4. Again, we obtain similar results indicating that long-memory FARIMA models are not preferred over short-memory ARMA models for the STL deseasoned Potsdam temperature data. This implies that trend or nonstationarity in the mean temperature does not play a critical role in the intensity of long memory in the Potsdam data.

FIG. 10. STL decomposition of the Potsdam temperature data in the period 1986–2015 into trend, seasonal, and remainder components (vertical scales are in $^{\circ}\text{C}$).

In [section 4](#), we operated on data deseasoned via Fourier notch filtering of the annual cycle and its first harmonic. This spectral filtering removes a seasonal cycle that is essentially constant from year to year, whereas STL removes a seasonal cycle that may vary from year to year. Since the removal of the STL seasonal cycle essentially removes long memory from the data, we conclude that interannual variability of the seasonal cycle plays a critical role in the presence of long memory in the Potsdam temperature data. Further analysis of teleconnection patterns, for example, will facilitate physical interpretations of the intriguing nonlinear seasonal cycle.

6. Discussion and relevance to earlier work

In this section, we establish the relevant context and compare our approach with earlier work. Recognition of the effect of serial correlation on the variability of climatic mean state estimates dates back to [Leith \(1973\)](#), who derived the variance of a finite time average of a first-order continuous-time autoregressive process in terms of its autocorrelation function. [Jones \(1975\)](#) and [Madden \(1979\)](#) extended Leith's results to discretely sampled red-noise climate processes and derived the variance of mean states in terms of the power spectral density.

Emerging from these inspiring works is the notion that the variance of a time average of autoregressive data is proportional to the variance of the time average that would be expected if data were independent. This proportionality factor, which depends only on the autocorrelation structure, was interpreted as the time between effectively independent samples ([Leith 1973](#); [Madden 1979](#)). Dividing the sample size used to compute the time average by this quantity results in what has been termed “effective sample size,” intended to represent the number of independent pieces of information in the data sequence ([Thiébaux and Zwiers 1984](#)). In [appendix A](#), we show that the approach presented in this paper generalizes this earlier approach to cases with both short- and long-range dependence.

With the ability to estimate the variability in time averages came the testing of hypotheses about differences in climate mean states. [Jones \(1975\)](#), [Katz \(1982\)](#), [Thiébaux and Zwiers \(1984\)](#), [Zwiers and Thiébaux \(1987\)](#), and [Zwiers and von Storch \(1995\)](#) provided various sorts of statistical hypothesis tests for detecting differences in climate mean states. Recently, the use of statistical hypothesis testing as the gold standard in research has come into question (e.g., [Nuzzo 2014](#); [Starbuck 2016](#)). Hence, in this work we use confidence intervals rather than hypothesis tests—following the lead of [Hayashi \(1982\)](#), who argued that confidence

intervals are more useful than a binary test result in the comparison of climate statistics.

With the widespread detection of long memory in climate data comes the need to extend the existing research on climate mean state variability, which focused solely on short-memory processes, to processes with long memory. Recently, [Massah and Kantz \(2016, hereafter MK16\)](#) described an approach for creating confidence intervals for time averages of processes with long-range correlations. Their approach involves manual tuning of the parameters of a FARIMA(1, d , 0) model and a graphical procedure for estimating the variance of mean states. Our approach improves on their methodology by fitting models with maximum likelihood, rather than manual tuning, and selecting the appropriate model order with an information criterion, rather than using only the FARIMA(1, d , 0) model. This is advantageous both in terms of reproducibility and computational feasibility, since maximum likelihood estimation and model selection can be automated, and in terms of accuracy, since underspecified FARIMA models can lead to biased parameter estimates ([Taqqu and Teverovsky 1998](#)). In addition, we provide an explicit formula [Eq. (5)] for the variance of time averages, which ameliorates the need to estimate variances graphically.

To aid in comparison of our approach with that of [MK16](#), we analyzed the same 123-yr dataset of daily average surface temperatures from Potsdam, Germany. [MK16](#) used the entire 123-yr record to calibrate their model and variance estimate, resulting in identical widths for the confidence intervals of the four 30-yr periods considered. However, if comparison among mean states is desired, and there is the possibility of climate change impacts from one period to another, then the statistics of each time period should be considered separately and not pooled together. While the graphical approach of [MK16](#) may require an extensive sample size to obtain reliable variance estimates, the parametric approach described in this paper does not require data beyond the period of interest, which allows for the separate estimation of the statistics for each 30-yr period in the Potsdam temperature record. The approach in this paper results in confidence intervals that are roughly consistent with those of [MK16](#), although those in [MK16](#) do not reflect the substantial increase in uncertainty during the most recent period: $\pm 0.34^{\circ}\text{C}$ (1896–1925), $\pm 0.48^{\circ}\text{C}$ (1926–55), $\pm 0.41^{\circ}\text{C}$ (1956–85), and $\pm 0.71^{\circ}\text{C}$ (1986–2015) (see [Table 4](#)) versus $\pm 0.5^{\circ}\text{C}$ for all periods in [MK16](#). Last, studies have sometimes found an additional long-memory weather regime at short time scales up to around two weeks, such as in atmospheric convection, clouds, and precipitation (e.g., [Tung et al. 2004, 2018](#)). The FARIMA models can still serve as crude approximations of such processes, with the short-memory part approximating the high-frequency

scaling regime and the long-memory part approximating the low-frequency regime.

7. Conclusions

This paper presents an approach for estimating variability and constructing confidence intervals for climate mean states, respecting both short- and long-range dependence. In particular, we make the following contributions:

- We demonstrate that *both* short- and long-range dependence structures in a temporal process must be considered to adequately characterize variability of mean states on a given time scale (section 2).
- We propose an adaptive and computationally feasible procedure for estimating the variability and constructing confidence intervals for the mean of climate data with both short- and long-range dependence (sections 3 and 4). The procedure is based on parametric modeling, selection among competing models using the Bayesian information criterion followed by the average–transform–smooth diagnostic visualization, and direct variance estimation from fitted model parameters.
- We use the proposed procedure and a dataset of 123 years of daily measurements to estimate the variability and determine error bars (confidence intervals) of each of four 30-yr mean states for the surface temperature at Potsdam, Germany (section 4). These confidence intervals are roughly twice the width as those obtained using prevailing methods, which disregard long memory. While the prevailing error bars assuming pure short memory indicate a significant increase in the mean temperature state in the most recent 30-yr period (1986–2015) relative to any of the three preceding 30-yr periods, the new error bars accounting for short and long memories indicate a significant change in mean temperature state only between the earliest (1896–1925) and the most recent period.

These contributions emphasize the fact that the width of confidence intervals or error bars bracketing estimated climate mean states depends critically on the dependence structure assumed for atmospheric variability. Moreover, the interannual variability of the seasonal cycle appears to be crucial for the long-memory property of the Potsdam temperature data and deserves further studies for physical interpretations. The approach can be customized and refined by incorporating physical information of a given climate regime in future applications. As evidence of long memory in climate data accumulates (e.g., Lovejoy and Schertzer 1986; Koscielny-Bunde et al. 1998; Huybers and Curry 2006; Vyushin and Kushner 2009; Franzke 2012; Yuan et al. 2015), representations of uncertainty for climate mean states should account for both short and

long memory and should certainly not assume pure short memory a priori. Hence, we recommend more meticulous consideration of the correlation structures of climate data—especially of their long-memory properties—in assessing the variability and determining confidence intervals for their mean states.

Acknowledgments. The authors are grateful to R. Sadikni at the University of Hamburg for the assistance to retrieve Potsdam data from the German Weather Service Climate Data Center. They thank Dr. W. S. Cleveland, Dr. D. Giannakis, and W. Huang for the stimulating discussions. MB's work has been supported by the Purdue Bilsland Fellowship and the NASA Earth and Space Science Fellowship under NASA-NNX16AO62H. MB and WT also acknowledge the support from the Purdue IMPACT program.

APPENDIX A

Generalization of Mean State Variance for Short-Memory Climate Processes

Here we show the formula for the variance of the mean of a short- and long-memory climate process in Eq. (5) generalizes that for a pure short-memory climate process given in Eqs. (10) and (11) of Katz (1982). Katz provides the variance for the sample mean $\bar{X}_n$ of an AR(p) process as

$$\text{Var}[\bar{X}_n] \sim \frac{V(\phi)}{n}, \quad n \rightarrow \infty, \quad (\text{A1})$$

where

$$V(\phi) = \frac{\sigma_\varepsilon^2}{\left(1 - \sum_{k=1}^p \phi_k\right)^2}. \quad (\text{A2})$$

Noting that the AR(p) process is FARIMA($p, 0, 0$) and using Eq. (6) for $f(n^{-1})$ we have

$$\begin{aligned} \text{Var}[\bar{X}_n] &\sim \nu(d)f(n^{-1})n^{-1} \\ &\sim 2\pi \left[\frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{\left(1 - \sum_{i=1}^p \phi_i\right)^2} \right] n^{-1}, \\ &\sim \frac{\sigma_\varepsilon^2}{n \left(1 - \sum_{k=1}^p \phi_k\right)^2} \end{aligned}$$

which is consistent with Eqs. (A1) and (A2).

TABLE B1. Outcomes of methods for choosing between the ARMA and FARIMA model for each 30-yr period. Results supporting the FARIMA model are in bold, results supporting the ARMA model are in italic, and neutral results are in regular typeface.

Period	ΔBIC	GOF p value		LRT p value	Rust (2007)	Chosen model
		ARMA	FARIMA			
1896–1925	<i>−0.9</i>	0.20	0.18	0.004	FARIMA	FARIMA
1926–55	15.5	0.94	0.97	<0.001	FARIMA	FARIMA
1956–85	3.9	0.36	0.47	—	Impartial	FARIMA
1986–2015	3.3	0.32	0.31	—	Impartial	FARIMA

APPENDIX B

Summative Model Selection Strategies

Several summative approaches for choosing between candidate ARMA and FARIMA models are available, including ΔBIC , goodness-of-fit testing, the likelihood-ratio test, and the simulation-based selection strategy of Rust (2007). We have already used the BIC to select the best model from each candidate pool, so it is convenient to use it to compare the best ARMA with the best FARIMA model. This can be done by considering the $\Delta\text{BIC} = \text{BIC}_{\text{ARMA}} - \text{BIC}_{\text{FARIMA}}$ (e.g., Raftery 1995; Kass and Raftery 1995). Since small BIC values are preferred, positive values of ΔBIC provide evidence supporting the FARIMA model, whereas negative values of ΔBIC support the ARMA model. Following the interpretation of Raftery (1995), evidence is weak for $0 < |\Delta\text{BIC}| < 2$, positive for $2 < |\Delta\text{BIC}| < 6$, strong for $6 < |\Delta\text{BIC}| < 10$, and very strong for $|\Delta\text{BIC}| > 10$.

Another option is to use the goodness-of-fit (GOF) test of Beran (1989). This procedure tests the null hypothesis that some sequence of time series data is generated by a given model against the alternative hypothesis that it is not. For a given time series, we can apply the GOF test using the best ARMA as the null model and then again using the best FARIMA as the null model. If both models are rejected or both are not rejected, the outcome is ambiguous. One model being rejected while the other is not rejected can be interpreted as evidence against the rejected model.

Another option is to compare the two models using a likelihood-ratio test (LRT), which tests the null hypothesis that the simpler model (ARMA) is an admissible simplification of the more complex model (FARIMA) against the alternative hypothesis that it is not (Neyman and Pearson 1933). A notable limitation of this method is that the simpler model must be nested in the more complex model (e.g., Kass and Raftery 1995). That is, it must be possible to obtain the simpler model by fixing the values of certain parameters in the more complex model; that is, the ARMA(0, 1) is nested in the

FARIMA(1, d , 1) because the ARMA(0, 1) can be obtained by fixing the FARIMA(1, d , 1) parameters d and ϕ to zero.

Finally, Rust (2007) developed a simulation-based model selection strategy for discriminating models in the FARIMA(p , d , q) class. Rust specifically uses this strategy to select between a candidate long-memory FARIMA model and a short-memory ARMA model. Given an observed sequence $\mathbf{x}$, consider the likelihood ratio $\text{lr}_{\text{obs}} = l_f(\hat{\Theta} | \mathbf{x}) - l_g(\hat{\Xi} | \mathbf{x})$, where $l_f(\hat{\Theta} | \mathbf{x})$ is the log-likelihood of model f with parameter estimate $\hat{\Theta}$, and $l_g(\hat{\Xi} | \mathbf{x})$ is the log-likelihood of model g with parameter estimate $\hat{\Xi}$. The idea is to compare lr_{obs} with the distributions of lr_f and lr_g (Hinde 1992), which are obtained analogously to lr_{obs} but from many realizations $\mathbf{x}_f$ and $\mathbf{x}_g$ simulated from models f and g , respectively. If the distributions lr_f and lr_g are well separated and lr_{obs} falls as a typical value from one distribution but a very rare value from the other, then this provides support for one model or the other. If on the other hand, the distributions lr_f and lr_g are not well separated then it may be difficult to distinguish them.

We can use these methods as a supplement to the diagnostic plot described in the main text (section 4) to choose between the best ARMA and best FARIMA models for each of the 30-yr periods in the Potsdam temperature data (Table B1). For the period 1896–1925, the ΔBIC between the ARMA(1, 2) and FARIMA(1, d , 2) is -0.9 (weak evidence favoring the ARMA), and the goodness-of-fit test (Beran 1992) fails to reject either model. However, a likelihood-ratio test rejects the hypothesis that the ARMA(1, 2) is an admissible simplification of the FARIMA(1, d , 2) (we use significance level $\alpha = 0.05$ throughout).

Figure B1 shows the empirical cumulative distribution functions (ECDFs) of $\text{lr}_{\text{FARIMA}}$ and lr_{ARMA} along with lr_{obs} for each 30-yr period. For the periods 1896–1925 and 1926–55, the distributions of $\text{lr}_{\text{FARIMA}}$ and lr_{ARMA} are well separated, with lr_{obs} falling as a typical value in the distribution of $\text{lr}_{\text{FARIMA}}$ and as a rare value in lr_{ARMA} , indicating the FARIMA models to be superior. The outcome is essentially the same for the period 1926–55,

FIG. B1. ECDFs of simulated likelihood ratios lr_{FARIMA} and lr_{ARMA} along with lr_{obs} (vertical dashed line) for each 30-yr period in the Potsdam temperature record.

except $\Delta\text{BIC} = 15.5$, providing very strong evidence in favor of the FARIMA(3, d , 0) over the ARMA(3, 0). Goodness-of-fit testing again fails to reject either model, and both the likelihood-ratio test and Rust's selection strategy clearly favor the FARIMA(3, d , 0).

For 1956–85, $\Delta\text{BIC} = 3.9$, which is positive evidence favoring the FARIMA(3, d , 0) over the ARMA(2, 2); the result for 1986–2015 is similar with $\Delta\text{BIC} = 3.3$, positive evidence favoring the FARIMA(3, d , 0) over the ARMA(2, 2). In both periods, the goodness-of-fit test again fails to reject either model. Rust's selection strategy also results in ambiguity with distributions that are not well separated with neither model clearly

preferred over the other. Since in both periods the two models are nonnested, we cannot use the likelihood-ratio test to inform a decision. Although in these two periods the testing procedures could not determine a preferred model, the ΔBIC does supply positive evidence for the FARIMA models. In this case, we choose the FARIMA models over the ARMA models in both periods 1956–85 and 1986–2015. Table B1 summarizes the results of these methods for choosing between the FARIMA and ARMA models and indicates that, based on these summative metrics, the FARIMA model is preferred over the ARMA model in all four 30-yr periods of the Potsdam temperature data.

REFERENCES

- Ailliot, P., and V. Monbet, 2012: Markov-switching autoregressive models for wind time series. *Environ. Modell. Software*, **30**, 92–101, <https://doi.org/10.1016/j.envsoft.2011.10.011>.
- Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723, <https://doi.org/10.1109/TAC.1974.1100705>.
- Ashkenazy, Y., D. R. Baker, H. Gildor, and S. Havlin, 2003: Nonlinearity and multifractality of climate change in the past 420,000 years. *Geophys. Res. Lett.*, **30**, 2146, <https://doi.org/10.1029/2003GL018099>.
- Beran, J., 1989: A test of location for data with slowly decaying serial correlations. *Biometrika*, **76**, 261–269, <https://doi.org/10.1093/biomet/76.2.261>.
- , 1992: A goodness-of-fit test for time series with long-range dependence. *J. Roy. Stat. Soc.*, **54B**, 749–760.
- , R. J. Bhansali, and D. Ocker, 1998: On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes. *Biometrika*, **85**, 921–934, <https://doi.org/10.1093/biomet/85.4.921>.
- , Y. Feng, S. Ghosh, and R. Kulik, 2013: *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Springer Berlin Heidelberg, 884 pp.
- Bowers, M. C., J. B. Gao, and W.-W. Tung, 2013: Long range correlations in tree ring chronologies of the USA: Variation within and across species. *Geophys. Res. Lett.*, **40**, 568–572, <https://doi.org/10.1029/2012GL054011>.
- Box, G. E. P., and G. M. Jenkins, 1970: *Time Series Analysis: Forecasting and Control*. Holden Day, 553 pp.
- Bunde, A., J. Ludescher, C. L. Franzke, and U. Büntgen, 2014: How significant is west Antarctic warming? *Nat. Geosci.*, **7**, 246–247, <https://doi.org/10.1038/ngeo2126>.
- Burnham, K. P., and D. R. Anderson, 2002: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, 488 pp.
- Caballero, R., S. Jewson, and A. Brix, 2002: Long memory in surface air temperature detection, modeling, and application to weather derivative valuation. *Climate Res.*, **21**, 127–140, <https://doi.org/10.3354/cr021127>.
- Chang, T. J., M. L. Kavvas, and J. W. Delleur, 1984: Daily precipitation modeling by discrete autoregressive moving average processes. *Water Resour. Res.*, **20**, 565–580, <https://doi.org/10.1029/WR020i005p00565>.
- Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning, 1990: STL: A seasonal-trend decomposition procedure based on Loess. *J. Off. Stat.*, **6**, 3–73.
- Cleveland, W. S., C. L. Mallows, and J. E. McRae, 1993: ATS methods: Nonparametric regression for non-Gaussian data. *J. Amer. Stat. Assoc.*, **88**, 821–835, <https://doi.org/10.1080/01621459.1993.10476347>.
- Eichner, J. F., E. Koscielny-Bunde, A. Bunde, S. Havlin, and H.-J. Schellnhuber, 2003: Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. *Phys. Rev. E*, **68**, 046133, <https://doi.org/10.1103/PhysRevE.68.046133>.
- Erdem, E., and J. Shi, 2011: ARMA based approaches for forecasting the tuple of wind speed and direction. *Appl. Energy*, **88**, 1405–1414, <https://doi.org/10.1016/j.apenergy.2010.10.031>.
- Fox, R., and M. S. Taqqu, 1986: Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Stat.*, **14**, 517–532, <https://doi.org/10.1214/aos/1176349936>.
- Franzke, C., 2010: Long-range dependence and climate noise characteristics of Antarctic temperature data. *J. Climate*, **23**, 6074–6081, <https://doi.org/10.1175/2010JCLI3654.1>.
- , 2012: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *J. Climate*, **25**, 4172–4183, <https://doi.org/10.1175/JCLI-D-11-00293.1>.
- , T. Graves, N. W. Watkins, R. B. Gramacy, and C. Hughes, 2012: Robustness of estimators of long-range dependence and self-similarity under non-Gaussianity. *Philos. Trans. Roy. Soc. London*, **370A**, 1250–1267, <https://doi.org/10.1098/rsta.2011.0349>.
- Gao, J., J. Hu, W.-W. Tung, Y. Cao, N. Sarshar, and V. P. Roychowdhury, 2006: Assessment of long-range correlation in time series: How to avoid pitfalls. *Phys. Rev. E*, **73**, 016117, <https://doi.org/10.1103/PhysRevE.73.016117>.
- , —, and —, 2011: Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLOS ONE*, **6**, e24331, <https://doi.org/10.1371/journal.pone.0024331>.
- Gil-Alana, L. A., 2005: Statistical modeling of the temperatures in the Northern Hemisphere using fractional integration techniques. *J. Climate*, **18**, 5357–5369, <https://doi.org/10.1175/JCLI3543.1>.
- Giraitis, L., and D. Surgailis, 1990: A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate. *Probab. Theory Relat. Fields*, **86**, 87–104, <https://doi.org/10.1007/BF01207515>.
- , P. Kokoszka, R. Leipus, and G. Teyssière, 2003: Rescaled variance and related tests for long memory in volatility and levels. *J. Econom.*, **112**, 265–294, [https://doi.org/10.1016/S0304-4076\(02\)00197-5](https://doi.org/10.1016/S0304-4076(02)00197-5).
- Granger, C. W., and R. Joyeux, 1980: An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Anal.*, **1**, 15–29, <https://doi.org/10.1111/j.1467-9892.1980.tb00297.x>.
- Graves, T., R. Gramacy, C. Franzke, and N. Watkins, 2015: Efficient Bayesian inference for natural time series using ARFIMA processes. *Nonlinear Processes Geophys.*, **22**, 679–700, <https://doi.org/10.5194/npgd-2-573-2015>.
- Gutzler, D. S., and K. C. Mo, 1983: Autocorrelation of Northern Hemisphere geopotential heights. *Mon. Wea. Rev.*, **111**, 155–164, [https://doi.org/10.1175/1520-0493\(1983\)111<0155:AONHGH>2.0.CO;2](https://doi.org/10.1175/1520-0493(1983)111<0155:AONHGH>2.0.CO;2).
- Hayashi, Y., 1982: Confidence intervals of a climatic signal. *J. Atmos. Sci.*, **39**, 1895–1905, [https://doi.org/10.1175/1520-0469\(1982\)039<1895:CIOACS>2.0.CO;2](https://doi.org/10.1175/1520-0469(1982)039<1895:CIOACS>2.0.CO;2).
- Hinde, J., 1992: Choosing between non-nested models: A simulation approach. *Advances in GLIM and Statistical Modelling*, L. Fahrmeir et al., Eds., Springer, 119–124.
- Hosking, J. R. M., 1981: Fractional differencing. *Biometrika*, **68**, 165–176, <https://doi.org/10.1093/biomet/68.1.165>.
- Hurst, H. E., 1951: Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civ. Eng.*, **116**, 770–808.
- Huybers, P., and W. Curry, 2006: Links between annual, Milankovitch and continuum temperature variability. *Nature*, **441**, 329–332, <https://doi.org/10.1038/nature04745>.
- Jolliffe, I. T., 1983: Quasi-periodic meteorological series and second-order autoregressive processes. *J. Climatol.*, **3**, 413–417, <https://doi.org/10.1002/joc.3370030409>.
- Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.*, **14**, 159–163, [https://doi.org/10.1175/1520-0450\(1975\)014<0159:ETVOTA>2.0.CO;2](https://doi.org/10.1175/1520-0450(1975)014<0159:ETVOTA>2.0.CO;2).

- Kantelhardt, J. W., E. Koscielny-Bunde, D. Rybski, P. Braun, A. Bunde, and S. Havlin, 2006: Long-term persistence and multifractality of precipitation and river runoff records. *J. Geophys. Res.*, **111**, D01106, <https://doi.org/10.1029/2005JD005881>.
- Kass, R. E., and A. E. Raftery, 1995: Bayes factors. *J. Amer. Stat. Assoc.*, **90**, 773–795, <https://doi.org/10.1080/01621459.1995.10476572>.
- Katz, R. W., 1982: Statistical evaluation of climate experiments with general circulation models: A parametric time series modeling approach. *J. Atmos. Sci.*, **39**, 1446–1455, [https://doi.org/10.1175/1520-0469\(1982\)039<1446:SECEW>2.0.CO;2](https://doi.org/10.1175/1520-0469(1982)039<1446:SECEW>2.0.CO;2).
- , and R. H. Skaggs, 1981: On the use of autoregressive-moving average processes to model meteorological time series. *Mon. Wea. Rev.*, **109**, 479–484, [https://doi.org/10.1175/1520-0493\(1981\)109<0479:OTUOAM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1981)109<0479:OTUOAM>2.0.CO;2).
- Kavasseri, R. G., and K. Seetharaman, 2009: Day-ahead wind speed forecasting using *f*-ARIMA models. *Renewable Energy*, **34**, 1388–1393, <https://doi.org/10.1016/j.renene.2008.09.006>.
- Klein, W. H., 1951: A hemispheric study of daily pressure variability at sea level and aloft. *J. Meteor.*, **8**, 332–346, [https://doi.org/10.1175/1520-0469\(1951\)008<0332:AHSDP>2.0.CO;2](https://doi.org/10.1175/1520-0469(1951)008<0332:AHSDP>2.0.CO;2).
- Koscielny-Bunde, E., A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, and H.-J. Schellnhuber, 1998: Indication of a universal persistence law governing atmospheric variability. *Phys. Rev. Lett.*, **81**, 729, <https://doi.org/10.1103/PhysRevLett.81.729>.
- Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin, 1992: Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econom.*, **54**, 159–178, [https://doi.org/10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y).
- Leith, C. E., 1973: The standard error of time-average estimates of climatic means. *J. Appl. Meteor.*, **12**, 1066–1069, [https://doi.org/10.1175/1520-0450\(1973\)012<1066:TSEOTA>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<1066:TSEOTA>2.0.CO;2).
- Lennartz, S., and A. Bunde, 2009: Trend evaluation in records with long-term memory: Application to global warming. *Geophys. Res. Lett.*, **36**, L16706, <https://doi.org/10.1029/2009GL039516>.
- , and —, 2011: Distribution of natural trends in long-term correlated records: A scaling approach. *Phys. Rev. E*, **84**, 021129, <https://doi.org/10.1103/PhysRevE.84.021129>.
- Lin, J. W.-B., and J. D. Neelin, 2002: Considerations for stochastic convective parameterization. *J. Atmos. Sci.*, **59**, 959–975, [https://doi.org/10.1175/1520-0469\(2002\)059<0959:CFSCP>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<0959:CFSCP>2.0.CO;2).
- Lindgren, B. W., 1976: *Statistical Theory*. MacMillan, 614 pp.
- Lovejoy, S., and B. B. Mandelbrot, 1985: Fractal properties of rain, and a fractal model. *Tellus*, **37A**, 209–232, <https://doi.org/10.1111/j.1600-0870.1985.tb00423.x>.
- , and D. Schertzer, 1986: Scale invariance in climatological temperatures and the local spectral plateau. *Ann. Geophys.*, **4B**, 401–410.
- , and —, 2012: Haar wavelets, fluctuations and structure functions: Convenient choices for geophysics. *Nonlinear Processes Geophys.*, **19**, 513–527, <https://doi.org/10.5194/npg-19-513-2012>.
- Ludescher, J., A. Bunde, C. L. Franzke, and H. J. Schellnhuber, 2016: Long-term persistence enhances uncertainty about anthropogenic warming of Antarctica. *Climate Dyn.*, **46**, 263–271, <https://doi.org/10.1007/s00382-015-2582-5>.
- Madden, R. A., 1976: Estimates of the natural variability of time-averaged sea-level pressure. *Mon. Wea. Rev.*, **104**, 942–952, [https://doi.org/10.1175/1520-0493\(1976\)104<0942:EOTNVO>2.0.CO;2](https://doi.org/10.1175/1520-0493(1976)104<0942:EOTNVO>2.0.CO;2).
- , 1979: A simple approximation for the variance of meteorological time averages. *J. Appl. Meteor.*, **18**, 703–706, [https://doi.org/10.1175/1520-0450\(1979\)018<0703:ASAFTV>2.0.CO;2](https://doi.org/10.1175/1520-0450(1979)018<0703:ASAFTV>2.0.CO;2).
- , and W. Sadeh, 1975: Empirical estimates of the standard error of time-averaged climatic means. *J. Appl. Meteor.*, **14**, 164–169, [https://doi.org/10.1175/1520-0450\(1975\)014<0164:EEOTSE>2.0.CO;2](https://doi.org/10.1175/1520-0450(1975)014<0164:EEOTSE>2.0.CO;2).
- , and D. J. Shea, 1978: Estimates of the natural variability of time-averaged temperatures over the United States. *Mon. Wea. Rev.*, **106**, 1695–1703, [https://doi.org/10.1175/1520-0493\(1978\)106<1695:EOTNVO>2.0.CO;2](https://doi.org/10.1175/1520-0493(1978)106<1695:EOTNVO>2.0.CO;2).
- Majda, A. J., and B. Gershgorin, 2010: Quantifying uncertainty in climate change science through empirical information theory. *Proc. Natl. Acad. Sci. USA*, **107**, 14 958–14 963, <https://doi.org/10.1073/pnas.1007009107>.
- Mandelbrot, B. B., and J. R. Wallis, 1968: Noah, Joseph, and operational hydrology. *Water Resour. Res.*, **4**, 909–918, <https://doi.org/10.1029/WR004i005p00909>.
- Massah, M., and H. Kantz, 2016: Confidence intervals for time averages in the presence of long-range correlations, a case study on Earth surface temperature anomalies. *Geophys. Res. Lett.*, **43**, 9243–9249, <https://doi.org/10.1002/2016GL069555>.
- Montanari, A., R. Rosso, and M. S. Taqqu, 1996: Some long-run properties of rainfall records in Italy. *J. Geophys. Res.*, **101**, 29 431–29 438, <https://doi.org/10.1029/96JD02512>.
- National Academies of Sciences, Engineering, and Medicine, 2016: *Next Generation Earth System Prediction: Strategies for Sub-seasonal to Seasonal Forecasts*. The National Academies Press, 350 pp., <https://doi.org/10.17226/21873>.
- Neyman, J., and E. Pearson, 1933: On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc.*, **231**, 289–337, <https://doi.org/10.1098/rsta.1933.0009>.
- Nuzzo, R., 2014: Statistical errors. *Nature*, **506**, 150–152, <https://doi.org/10.1038/506150a>.
- Pandey, G., S. Lovejoy, and D. Schertzer, 1998: Multifractal analysis including extremes of daily river flow series for basins one to a million square kilometers. *J. Hydrol.*, **208**, 62–81, [https://doi.org/10.1016/S0022-1694\(98\)00148-6](https://doi.org/10.1016/S0022-1694(98)00148-6).
- Pelletier, J. D., 1998: The power spectral density of atmospheric temperature from time scales of 10^2 to 10^6 yr. *Earth Planet. Sci. Lett.*, **158**, 157–164, [https://doi.org/10.1016/S0012-821X\(98\)00051-X](https://doi.org/10.1016/S0012-821X(98)00051-X).
- , and D. L. Turcotte, 1997: Long-range persistence in climatological and hydrological time series: Analysis, modeling and application to drought hazard assessment. *J. Hydrol.*, **203**, 198–208, [https://doi.org/10.1016/S0022-1694\(97\)00102-9](https://doi.org/10.1016/S0022-1694(97)00102-9).
- Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, 1994: Mosaic organization of DNA nucleotides. *Phys. Rev. E*, **49**, 1685, <https://doi.org/10.1103/PhysRevE.49.1685>.
- Raftery, A. E., 1995: Bayesian model selection in social research. *Sociol. Methodol.*, **25**, 111–163, <https://doi.org/10.2307/271063>.
- Rust, H. W., 2007: Detection of long-range dependence: Applications in climatology and hydrology. Ph.D. thesis, University of Potsdam, 165 pp., <http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-13347>.
- Santer, B., J. Hnilo, T. Wigley, J. Boyle, C. Doutriaux, M. Fiorino, D. Parker, and K. Taylor, 1999: Uncertainties in observationally based estimates of temperature change in the free atmosphere. *J. Geophys. Res.*, **104**, 6305–6333, <https://doi.org/10.1029/1998JD200096>.
- Schmitt, F., S. Lovejoy, and D. Schertzer, 1995: Multifractal analysis of the Greenland Ice-Core Project climate data. *Geophys. Res. Lett.*, **22**, 1689–1692, <https://doi.org/10.1029/95GL01522>.
- Schwarz, G., and Coauthors, 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, <https://doi.org/10.1214/aos/1176344136>.

- Shackleton, N. J., and J. Imbrie, 1990: The $\delta^{18}\text{O}$ spectrum of oceanic deep water over a five-decade band. *Climatic Change*, **16**, 217–230, <https://doi.org/10.1007/BF00134658>.
- Starbuck, W. H., 2016: 60th anniversary essay: How journals could improve research practices in social science. *Admin. Sci. Quart.*, **61**, 165–183, <https://doi.org/10.1177/0001839216629644>.
- Taqqu, M. S., and V. Teverovsky, 1998: On estimating the intensity of long-range dependence in finite and infinite variance time series. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, R. Adler, R. Feldman, and M. Taqqu, Eds., Birkhäuser, 177–217.
- Thiébaux, H. J., and F. W. Zwiers, 1984: The interpretation and estimation of effective sample size. *J. Climate Appl. Meteor.*, **23**, 800–811, [https://doi.org/10.1175/1520-0450\(1984\)023<0800:TIAEOE>2.0.CO;2](https://doi.org/10.1175/1520-0450(1984)023<0800:TIAEOE>2.0.CO;2).
- Thomas, E. R., P. F. Dennis, T. J. Bracegirdle, and C. Franzke, 2009: Ice core evidence for significant 100-year regional warming on the Antarctic Peninsula. *Geophys. Res. Lett.*, **36**, L20704, <https://doi.org/10.1029/2009GL040104>.
- Tsonis, A. A., P. J. Roebber, and J. B. Elsner, 1999: Long-range correlations in the extratropical atmospheric circulation: Origins and implications. *J. Climate*, **12**, 1534–1541, [https://doi.org/10.1175/1520-0442\(1999\)012<1534:LRCITE>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<1534:LRCITE>2.0.CO;2).
- Tung, W.-W., M. W. Moncrieff, and J.-B. Gao, 2004: A systemic analysis of multiscale deep convective variability over the tropical Pacific. *J. Climate*, **17**, 2736–2751, [https://doi.org/10.1175/1520-0442\(2004\)017<2736:ASAOMD>2.0.CO;2](https://doi.org/10.1175/1520-0442(2004)017<2736:ASAOMD>2.0.CO;2).
- , A. Barthur, M. C. Bowers, Y. Song, J. Gerth, and W. S. Cleveland, 2018: Divide and Recombine (D&R) data science projects for deep analysis of big data and high computational complexity. *Japanese J. Stat. Data Sci.*, <https://doi.org/10.1007/s42081-018-0008-4>, in press.
- von Storch, J.-S., P. Müller, and E. Bauer, 2001: Climate variability in millennium integrations with coupled atmosphere-ocean GCMs: A spectral view. *Climate Dyn.*, **17**, 375–389, <https://doi.org/10.1007/s003820000110>.
- Vyushin, D. I., and P. J. Kushner, 2009: Power-law and long-memory characteristics of the atmospheric general circulation. *J. Climate*, **22**, 2890–2904, <https://doi.org/10.1175/2008JCLI2528.1>.
- Weber, R. O., and P. Talkner, 2001: Spectra and correlations of climate data from days to decades. *J. Geophys. Res.*, **106**, 20 131–20 144, <https://doi.org/10.1029/2001JD000548>.
- Whittle, P., 1953: Estimation and information in stationary time series. *Ark. Mat.*, **2**, 423–434, <https://doi.org/10.1007/BF02590998>.
- World Meteorological Organization, 2011: Guide to climatological practices. Tech. Rep. WMO-100, 117 pp., http://www.wmo.int/pages/prog/wcp/ccl/guide/documents/WMO_100_en.pdf.
- Yuan, N., M. Ding, Y. Huang, Z. Fu, E. Xoplaki, and J. Luterbacher, 2015: On the long-term climate memory in the surface air temperature records over Antarctica: A non-negligible factor for trend evaluation. *J. Climate*, **28**, 5922–5934, <https://doi.org/10.1175/JCLI-D-14-00733.1>.
- Zwiers, F. W., and H. J. Thiébaux, 1987: Statistical considerations for climate experiments. Part I: Scalar tests. *J. Climate Appl. Meteor.*, **26**, 464–476, [https://doi.org/10.1175/1520-0450\(1987\)026<0464:SCFCEP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1987)026<0464:SCFCEP>2.0.CO;2).
- , and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336–351, [https://doi.org/10.1175/1520-0442\(1995\)008<0336:TSCIAI>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2).