12-2017

Variability of Persistent Temporal Correlation in Climate Data

Matthew C. Bowers
Purdue University

VARIABILITY OF PERSISTENT TEMPORAL CORRELATION IN CLIMATE DATA

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Matthew C. Bowers

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2017

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Alexander Gluhovsky, Chair

Department of Earth, Atmospheric, and Planetary Sciences

Dr. Wen-wen Tung

Department of Earth, Atmospheric, and Planetary Sciences

Dr. Wen-yih Sun

Department of Earth, Atmospheric, and Planetary Sciences

Dr. William S. Cleveland

Department of Statistics

Dr. Dimitris Giannakis

Courant Institute of Mathematical Sciences, New York University

Approved by:

Dr. Indrajeet Chaubey

Head, Department of Earth, Atmospheric, and Planetary Sciences

to Shiva
who stayed by the door to wait
even when I had to work late

ACKNOWLEDGMENTS

I feel tremendous gratitude to Prof. Wen-wen Tung for her mentorship and guidance throughout the epic journey that is graduate school. I am grateful to my colleagues Dr. Qianwen Luo and Dr. Wanchen Wu, who patiently answered all of my naive questions about meteorology. I am thankful for the chance to work with Dr. Jianbo Gao, who inspired me with his sense of creativity. I have been privileged to benefit from discussions with talented researchers including Dr. William S. Cleveland, Dr. Dimitris Giannakis, Dr. Alexander Gluhovsky, Dr. John Cushman, Dr. Wen-yih Sun, Dr. Bo Li, Dr. Philippe Gaudreau, and Dr. Whitney Huang. I am grateful to R. Sadikni at the University of Hamburg for the assistance retrieving Potsdam data from the German Weather Service Climate Data Center. I have been supported by the Purdue Bilsland Fellowship, the Purdue IMPACT program, NSF grants CMMI-1031958 and 0826119, and the NASA Earth and Space Science Fellowship under NASA-NNX16AO62H.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Bowers, Matthew C. PhD, Purdue University, December 2017. Variability of Persistent Temporal Correlation in Climate Data. Major Professor: Wen-wen Tung.

This dissertation examines manifestations of persistent memory in climate data. Persistence is characterized by a slow power-law decay in the autocorrelations of a time series. Its existence implies that the influence of past values in a time series extend into the distant future. It has numerous theoretical implications, notably that it changes the asymptotic decay in the variance of sample means, which can substantially impact the uncertainty in climate mean states. Its intensity can vary over space, time, and other dimensions, e.g. tree species. Variation in its intensity can be used for practical applications such as discriminating between steady and intermittent rainfall and assessing the calibration period needed for paleoclimate proxy data.

This work explores three major areas in which persistence can be leveraged to better understand the complexities of climate data. The first is in tree ring width data, which are among the best proxies for reconstructing paleoclimate records. The persistent correlations found in tree ring data suggest that the behavior of tree ring growth observed in a short calibration period may be similar to the general behavior of tree ring growth in a much longer period; therefore, the limited calibration period may be more useful than previously thought. The second area is in the quantification of uncertainty in the mean states of climate data. A framework for quantifying uncertainty in climate means is presented which can account for both classical short-range correlations and long-term persistent correlations. The final area is in the detection of subtle changes in tropical rainfall patterns. Persistence is used to illuminate recent changes in the temporal clustering patterns of rainfall in the tropical belt; the

detected changes could have critical implications for the water resource management of the affected regions.

1 LONG-RANGE CORRELATION IN TREE RING CHRONOLOGIES OF THE USA: VARIATION WITHIN AND ACROSS SPECIES

1.1 Introduction

Tree ring width data are abundant and have high temporal resolution. They respond well to environmental conditions such as solar radiation, air temperature, soil temperature, precipitation, soil moisture, and humidity [1,2]. Therefore, they are considered among the best proxies for reconstructing past climate records, and thus have been utilized for a wide variety of applications in paleoclimatology, including reconstruction of past temperature records [3–5] and past precipitation records [6, 7].

It is well-known that different species of tree respond differently to certain environmental conditions, e.g., some are more sensitive to temperature variation while others are more sensitive to moisture variation [1]. To study such sensitivities, one may examine how tree ring annual growth width or density correlates with various environmental factors over a certain calibration period [8]. Calibration periods are typically much shorter than the periods over which paleoclimate records are reconstructed, which are often on the scale of 10^3 years [3, 9]. Will problems arise due to limited calibration periods? To gain insights into this issue, we focus on studying scaling and the correlation structure of tree ring time series.

Noticing that temperature, rainfall, and river discharge time series may be reconstructed from tree ring data through simple linear regressions, it may be expected that tree ring data also exhibit the long-range correlations observed in temperature [10], rainfall [11], and streamflow [12]. Indeed, long-range correlation has been found in the raw tree ring width time series of two species [13, 14]. This motivates us to systematically examine the long-range correlation properties of tree ring chronologies

that cannot be readily captured by low-order auto-regressive modeling. In particular, we will examine how these properties may vary within and across tree species.

The long-range correlation properties of a time series can be conveniently characterized by the Hurst exponent or Hurst parameter H [15]. The Hurst parameter quantifies the persistence of a correlation such that when $0 < H < 1/2$, the signal has anti-persistent correlations; when $H = 1/2$, the signal is memoryless or has short-range correlations; when $1/2 < H < 1$, the signal has persistent long-range correlations; and when $H > 1$ the signal may be non-stationary or have non-trivial trends. The Hurst parameter can be estimated by a variety of methods; however, few of these methods are capable of accurately estimating H when $H > 1$ [16, 17]. Therefore care must be taken when estimating H from tree ring data, since H can often be larger than 1 [14].

One of the few methods capable of accurately estimating H when $H > 1$ is the popular detrended fluctuation analysis (DFA) [18]. The problem of estimating H when $H > 1$ can also be aptly dealt with by a more recent method called adaptive fractal analysis (AFA) [19], which performs comparably to DFA in many situations, but may be able to better deal with arbitrary trends in the signal [19]. Here these methods are used to characterize the long-range correlation properties of 10 different tree species represented in a database of 697 tree ring chronologies from sites across the continental United States.

1.2 Data

All of the tree ring site chronologies studied here were obtained from the NOAA Paleoclimatology Program's International Tree Ring Data Bank (ITRDB) at http://www.ncdc.noaa. For consistency, we use only chronologies created from tree ring width samples, excluding those from maximum latewood density. The ITRDB has over 1000 site chronologies from the United States, but since our purpose is to make statistical comparisons among tree species, we narrow the database to species which have at least 25 site

chronologies. This leaves representatives from 10 tree species at 697 sites; locations are mapped in Fig. 1.1, while the ITRDB tree species codes with their corresponding Latin and common names are listed in Table 1.1.

Tree ring chronologies are derived as follows. Beginning with multiple core samples from multiple trees at a given site, raw ring width measurements are combined using a process described in [20]– for convenience, it is summarized in Sec. 1 of the supplementary material. This procedure is intended to remove the effects of factors other than the controlling environment, such as growth trends in individual trees and intra- and inter-tree variability. This leaves one time series of growth indices per site, intended to serve as a proxy environmental signal capable of reflecting local climatic conditions.

1.3 Methodology

As pointed out earlier, the Hurst parameter H characterizes the long-range correlations in a time series. There are many effective ways to estimate H [16, 17]. Since H for tree ring width data may have $H > 1$ [14], while most of the methods available, such as fluctuation analysis and rescaled range (R/S) analysis, yield estimates of H that saturate at 1 [16, 17], care must be taken to choose the appropriate methods. Here, we choose detrended fluctuation analysis (DFA) [18], which is the most popular and is not hindered by the saturation problem. We also employ a newer method, adaptive fractal analysis (AFA) [19], which is comparable to DFA in many situations, but may better deal with arbitrary trends in the signals [19].

DFA works as follows: given a noise (or increment) time series, $x_1, x_2, x_3, \cdots$, with mean $\overline{x}$, one first constructs a random walk process,

$$u(i) = \sum_{k=1}^{i} (x_k - \overline{x}), \quad i = 1, 2, \cdots, N \tag{1.1}$$

One then divides $\{u(i),\ i = 1, 2, \cdots, N\}$ into N/l non-overlapping segments (where N/l denotes the largest integer equal to or smaller than N/l), each containing l points, and defines the local trend in each segment to be the ordinate of a best

linear or polynomial fit of the time series in that segment. Finally, one computes the "detrended walk", denoted by $u_l(n)$, as the difference between the original "walk" $u(n)$ and the local trend. The fractal behavior is described by the following scaling law

$$F_d(l) = \left\langle \sum_{i=1}^{l} u_l(i)^2 \right\rangle^{1/2} \sim l^H \tag{1.2}$$

where the angle brackets denote ensemble averages of all the segments. The so-called DFA-m denotes the use of an m-order polynomial for the segment fitting. For tree ring data, DFA-1 has been found to be consistent with its higher-order counterparts [14], so in this work, we focus on DFA-1.

AFA first estimates a globally smooth trend signal $v(i)$ for the random walk process $u(i)$, and thus eliminates the problem of abrupt jumps at the boundaries of neighboring segments in DFA. The residual, $u(i) - v(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to [19]

$$F(w) = \left\langle \frac{1}{N} \sum_{i=1}^{N} (u(i) - v(i))^2 \right\rangle^{1/2} \sim w^H. \tag{1.3}$$

1.4 Analysis and Results

1.4.1 Estimation of Hurst parameter

Based on the log-log plot of $F(w)$ versus w for each tree ring chronology, linear fitting is used to estimate the slope across the power-law scaling region, providing an estimate of H. Examples of this procedure are shown for four site chronologies in Fig. 1.2. The estimates of H are consistent between DFA-1 and AFA, thus the estimates obtained from AFA are used in further analysis.

It is found that for some chronologies studied here, the scaling behavior may only be defined for time scales up to about 33 years, corresponding to the first 5 points of AFA (see Fig. 1.2d), while for other chronologies, the scaling behavior may be defined for a few hundred years (see Fig. 1.2a,b,c). For ring data with long scaling regions,

the Hurst parameter determined by 5 points in AFA is similar to that determined by the entire scaling region; thus, in the following analysis, we will only use the first 5 points in AFA to estimate the Hurst parameter.

Careful examination of the data with shorter scaling ranges reveals that the break in scaling is due to a large cutoff in low frequency variation in the ring data. This is confirmed by the power spectral densities shown in Fig. 1.3. In particular, note the scaling break in the Jack's Fork chronology in Fig. 1.2d and the corresponding suppression of its low frequency variation in Fig. 1.3d. It is known that the processing of raw tree ring data can lead to loss of low frequency variation in the extracted chronologies [21, 22]. This suggests that the limited scaling in some chronologies may not be intrinsic to the data, but due to excessive filtering.

1.4.2 Variation of Hurst parameter within and across species

We first checked the variation of H within species. This is best illustrated by estimating the distribution for H for each species using a kernal density estimation method. Four examples are shown in Fig. 1.4. Alternatively, we may use scatter plots as shown in Fig. 1.5. Based on these results, we can compute sample statistics for the groups, which are given in Table 1.2. The mean value of H for each species population satisfies $1/2 < H < 1$, indicating long-memory and persistent behavior.

From Table 2 and Fig. 1.5, we also note that there is considerable variation in the mean H value from species to species, suggesting variation in the typical strength of persistence. To quantitatively examine variations of H across species, two procedures are adopted here. One is ANOVA, which tests the hypothesis of equal group means against the alternative of non-equality. The test, which is detailed in the Appendix, yields a p-value .0001, rejecting the hypothesis of equal group means. This result prompts the use of a multiple comparison procedure, also detailed in the Appendix, to further clarify the variation among species. The results of this analysis, with significance level $\alpha = .05$, are summarized in Table 1.2.

While it is desirable to find the connections between the long-range correlation properties of tree ring chronologies and climate variability, our exploratory analysis does not find any simple connections, as shown in Sec. 2 of the supplementary material. Due to potentially large variations in H over short inter-site distances, the most appropriate use of chronology correlation properties may be in spatial aggregations of H to assist with continental-, hemispheric-, or global-scale climate variable reconstructions.

1.5 Discussions and Conclusions

A database of 697 tree ring chronologies from 10 different tree species across the United States is examined for long-range correlation properties using detrended fluctuation analysis and adaptive fractal analysis. The Hurst parameter H is used to quantify the strength of these correlations. The mean value of H for sites of each species lies in $1/2 < H < 1$, indicating a tendency for persistent behavior. There is variation in the mean H across species within this interval, and analysis of variance and a multiple comparison procedure are used to detect differences in the mean H value among species. The difference in mean Hurst parameter value between certain species is found to be statistically significant.

Our study has two interesting implications. One concerns the usefulness of a limited calibration period, which is often short compared to the time span over which paleoclimatic variables are to be reconstructed. Had tree ring data only exhibited short-range correlations, short calibration periods would indeed have very limited value; fortunately, this is not the case — the self-similarity and long-range correlations in tree ring data implies that the general behavior of tree ring growth would be similar to that observed in the calibration period.

The second implication of our work concerns the reconstruction of paleoclimatic records. One critical question is: how may the tree ring data of different sites of the same species and different species be combined to best reconstruct paleoclimatic

records? An intelligent response might be to utilize the value of the Hurst parameters of each site to construct a suitable weighting scheme for sites of the same species, as well as for different species. Development of such a scheme would be an important task in future research.

1.6 Appendix: ANOVA and Multiple Comparison Procedure

ANOVA basically compares two types of variations – within groups and between groups, by assuming [23]:

1. Observations are independent
2. Group populations are normally distributed
3. Group populations have the same variance

To fix the idea, let us consider p groups, where the i-th group having observations $y_{ij}, j = 1, \cdots, n_i$, and mean $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$. Clearly, the global mean is $\overline{y} = \frac{1}{n} \sum_{i=1}^{p} n_i \overline{y}_i$, where $n = \sum_{i=1}^{p} n_i$. The total variation around the global mean may be written as

$$SS_{\text{Total}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2 = SS_{\text{Within}} + SS_{\text{Between}} \tag{1.4}$$

where

$$SS_{\text{Within}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 \tag{1.5}$$

$$SS_{\text{Between}} = \sum_{i=1}^{p} n_i (\overline{y}_i - \overline{y})^2 \tag{1.6}$$

The test statistic is obtained from the ratio

$$F = \frac{SS_{\text{Between}}/(p-1)}{SS_{\text{Within}}/(n-p)} \tag{1.7}$$

which follows an F-distribution with $p-1$ and $n-p$ degrees of freedom. When this null hypothesis is not true, the expectation of the ratio will be larger than it would

be if the null hypothesis were true. ANOVA gives a criterion for accepting/rejecting the null hypothesis for an observed F under the three basic assumptions. Note that in practice, assumptions (2) and (3) may be weakly violated.

ANOVA is most conveniently carried out in the free software R. In the following examples, '`>`' indicates a prompt for commands in the R environment. Begin by reading the data from a file `H_species.dat`.

```
> rings <- read.table("H_species.dat",header=TRUE)
```

Now the dataframe `rings` has two columns: H (observations) and species (groups). We can execute the ANOVA test and view the results.

```
> aov.rings <- aov(H~species,data=rings)
> summary(aov.rings)
```

The results of this one-way ANOVA test for differences in group mean H are given in Table 1.3. The small p-value indicates that the hypothesis of equal group means is rejected at any reasonable significance level.

Now that ANOVA has indicated differences among the groups, we need to determine whether two tree chronologies have the same mean Hurst parameter or not. Intuitively, when the difference between the means of two groups of observations is large compared with the summation of the two standard deviations, the groups may be considered to have different means. When there are only two groups, taking into account the effect of sample size, this idea leads to the t-test. When there are more than two groups, a procedure, called Tukey HSD (honestly significant difference) multiple comparison, can fix the probability α of falsely identifying differences in group means, for the entire set of pairwise comparisons, rather than for each pair individually. This reduces the overall chance of incorrectly rejecting the null hypotheses, and therefore, Tukey HSD multiple comparison is more suitable than a simple sequence of pairwise t-tests.

We can carry out the Tukey HSD procedure with $\alpha = .05$ in R as follows.

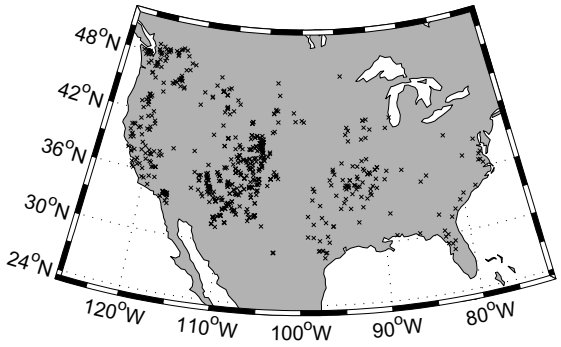

Figure 1.1. Map showing locations of site chronologies used in the study.

```
> tuk <- TukeyHSD(aov.rings)
> print(tuk)
```

The resulting R output gives a 95% confidence interval around the difference between each pair of group means. If the interval does not contain 0, then the true difference in group means is likely non-zero, and we consider the difference significant. For example, the first row of the output table is:

```
                diff          lwr        upr     p adj
PIED-QUDG 0.001274444 -0.056677522 0.05922641 1.0000000
```

where `diff` is the difference between group means of PIED and QUDG, `lwr` and `upr` are respectively the lower and upper confidence limits for the difference in group means, and `p adj` is the p-value after adjustment for the multiple comparisons. Here we see that the confidence interval about the difference in means between groups PIED and QUDG contains 0. Thus, there is not a significant difference in the two group means.

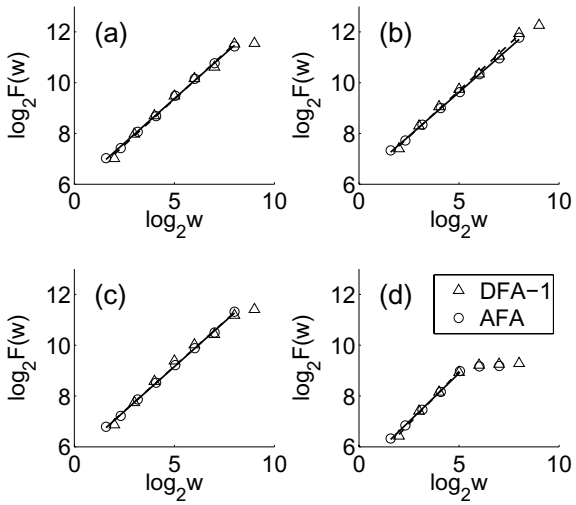

Figure 1.2. Long-range correlation analysis of tree ring chronologies from (a) Kane Spring, Utah (PIED), (b) Walnut Canyon National Monument, Arizona (PIPO), (c) Spruce Canyon, Colorado (PSME), and (d) Jack's Fork, Missouri (QUST). Circles indicate adaptive fractal analysis (AFA), while triangles indicate detrended fluctuation analysis of order 1 (DFA-1). AFA and DFA-1 are in good agreement for each time series.

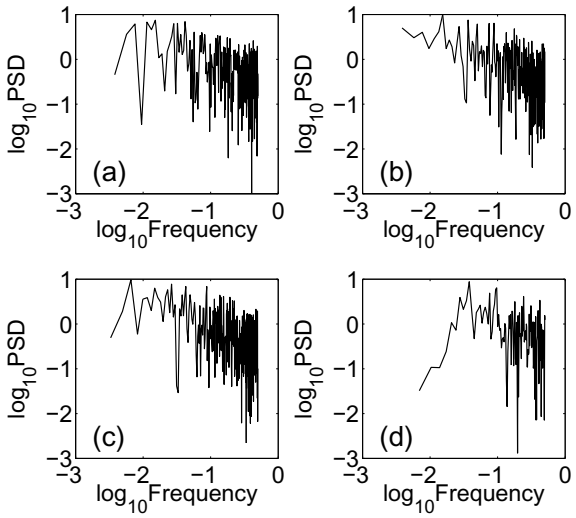

Figure 1.3. Power spectral density analysis of tree ring chronologies from (a) Kane Spring, Utah (PIED), (b) Walnut Canyon National Monument, Arizona (PIPO), (c) Spruce Canyon, Colorado (PSME), and (d) Jack's Fork, Missouri (QUST).

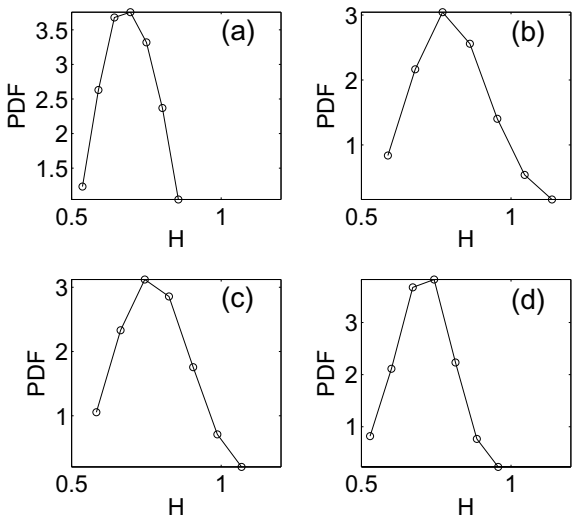

Figure 1.4. Probability density functions (PDF) of Hurst parameter for tree ring chronologies of (a) Pinyon pine (PIED), (b) Ponderosa pine (PIPO), (c) Douglas fir (PSME), and (d) post oak (QUST). Densities are obtained from kernal PDF estimation.

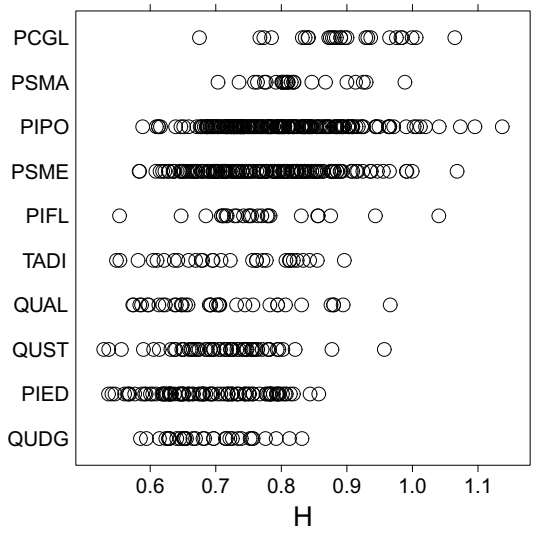

Figure 1.5. One-dimensional scatter plots of H by species

Table 1.1.
Tree species information

Species Code	Latin Name	Common Name
PCGL	*Picea glauca*	Canadian spruce
PIED	*Pinus edulis*	Colorado pinyon
PIFL	*Pinus flexilis*	limber pine
PIPO	*Pinus ponderosa*	ponderosa pine
PSMA	*Pseudotsuga macrocarpa*	bigcone Douglas-fir
PSME	*Pseudotsuga menziesii*	Douglas-fir
QUAL	*Quercus alba*	American white oak
QUDG	*Quercus douglasii*	blue oak
QUST	*Quercus stellata*	American post oak
TADI	*Taxodium distichum*	baldcypress

Table 1.2.
Summary statistics of H for each species. "Similar to" indicates which group means are not significantly different

Species Code	Similar to	Mean	Standard Deviation	Sample Size
[1]QUDG	2,3,4,5	0.690	0.063	33
[2]PIED	1,3,4,5	0.691	0.079	90
[3]QUST	1,2,4,5,6	0.709	0.077	57
[4]QUAL	1,2,3,5,6	0.711	0.105	35
[5]TADI	1,2,3,4,6,7	0.721	0.096	30
[6]PIFL	3,4,5,7,8,9	0.767	0.096	25
[7]PSME	5,6,8,9	0.773	0.093	168
[8]PIPO	6,7,9	0.800	0.096	207
[9]PSMA	6,7,8,10	0.823	0.066	25
[10]PCGL	9	0.892	0.085	27

Table 1.3.
One-way ANOVA test comparing mean values of H among the 10 species. "Groups" indicates variation between groups, while "Error" indicates variation within groups.

Source	Sum of Squares	df	Mean Squares	F	Prob $> F$
Groups	1.7880	9	0.1987	24.6836	.0001
Error	5.5294	687	0.0080		
Total	7.3174	696			

2 VARIABILITY AND CONFIDENCE INTERVALS FOR THE MEAN OF CLIMATE DATA WITH SHORT- AND LONG-RANGE DEPENDENCE

2.1 Introduction

A time average is perhaps the most fundamental way to characterize climate, but the average or mean of climate data during some time interval 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 [24], [25], and [26]. 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 [27–29].

However, another sort of temporal dependence structure called long-range dependence or long memory has since been detected in numerous physical state variables, e.g., rainfall [30, 31], streamflow [32–34], tropical deep convection [35], general circulation [36,37], surface temperature [37–42], and even climate proxies like ice cores [43] and tree rings [44]. The long memory implies a slow power-law decay in a process's autocorrelation 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 (Sec. 2.2). 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 (Sec. 2.3). 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-year time averages of the surface temperature at Potsdam, Germany (Sec. 2.4). We provide evidence that interannual variability of the seasonal cycle is a source of long memory in the Potsdam temperature data (Sec. 2.5). Discussions and comparison with related work are in Sec. 2.6, and concluding remarks are given in Sec. 2.7.

2.2 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.

2.2.1 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 ($t = 0, \ldots, N$), mean μ, and variance σ^2, the correlation structure is naturally expressed by the autocorrelation function

$$R(k) = \frac{\mathrm{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2} \tag{2.1}$$

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 [24, 28, 29]. 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, e.g., sea-level pressure [27, 45], surface temperature [46], and geopotential height [47].

More recently, evidence of long memory in various meteorological variables has also begun to accumulate, e.g., precipitation [30,31], tropical deep convection [35], general circulation [36, 37], and especially surface temperature [37–42, 48–51]. A stationary long-memory process exhibits power-law decay in its autocorrelation function, i.e., $R(k) \sim k^{2d-1}$ (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($\sum_{k=-\infty}^{\infty} R(k) = \infty$) [52]. 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 takes the form

$$f(\lambda) \sim c_f \lambda^{-2d} \quad (\lambda \to 0) \tag{2.2}$$

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 [32] by $H = d + 1/2$.

To illustrate these properties in 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 meters. The dataset, obtained from the German Weather Service Climate Data Center (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.

Figure 2.1 (left column) shows the time series of the raw Potsdam temperature data over the four 30-year periods: 1896–1925, 1926–1955, 1956-1985, 1986–2015. The data exhibit a prevalent seasonal cycle which dominates the dependence structure as seen in their autocorrelation function (Fig. 2.2a). 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-year time span similar to that in [28] in which observed data are considered as the superposition of a constant mean state, a seasonal cycle, potential non-stationary trends, and the remaining stochastic noise process. In addition to seasonality, if the data include significant non-stationary 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 [40, 53–55]. Adopting the methods of [40], we do not find significant trends in the Potsdam data and therefore do not detrend the data (see Sec. 2.4 for details). We remove the seasonal cycle by forward-backward notch filtering the annual and semiannual cycles, obtaining the temperature anomaly series shown in Fig. 2.1 (right column).

The autocorrelation function of these temperature anomalies, shown in Fig. 2.2c, 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. 2.2b and 2.2d, respectively) appear to exhibit power-law scaling at low frequencies, also consistent with long memory. Despite these indications of long memory, the temperature anomaly autocorrelations in the short-range up to a two week lag (not shown) exhibit exponential decay, consistent with short-memory. With the fingerprints 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.

2.2.2 Fractional ARIMA models

We now introduce the class of fractional autoregressive integrated moving-average (FARIMA) time series models, which can exhibit both short and long memories. This class is an extension of the classic short-memory ARMA models introduced by [56]. The ARMA class has been used in various meteorological and climatological applications including prediction of drought indices [57], modeling quasi-periodicity in zonal

circulation [58], modeling precipitation occurrence and magnitude [59], stochastic parameterization of atmospheric convective heating [60], and wind forecasting [61, 62]. The ARMA class models the values of a temporal process X_t sampled uniformly in time ($t = 1, \ldots, N$) as a linear combination of past values plus the linear combination of current and previous stochastic noise innovations, i.e.,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \ _t + \sum_{j=1}^{q} \theta_j \ _{t-j}, \quad t = 1, \ldots, N \tag{2.3}$$

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

[56] 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, [56] used a **backshift operator** B which imparts a lag to a temporal process, i.e., $B^k X_t = X_{t-k}$ 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).

[63] and [64] 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 non-stationary ARIMA processes. A fractional autoregressive integrated moving average, FARIMA(p, d, q), process Y_t is defined recursively by $Y_0 = 0$ and

$$(1 - B)^d Y_t = X_t \tag{2.4}$$

where X_t is an ARMA(p, q) process as defined by Eq. (2.3) and the differencing order d is a real number, identical to the long-memory parameter d discussed in Sec. 2.22.2.1. 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 $\theta_1, \theta_2, \ldots, \theta_q$ and the p autoregressive coefficients $\phi_1, \phi_2, \ldots, \phi_p$ (Eq. 2.3).

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, e.g., characterizing properties of rainfall [65], describing deep ocean variability [66], modeling surface temperatures [40,49], and forecasting wind speed [67].

2.2.3 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 $(t = 1, 2, \ldots, N)$ 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, i.e., 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 $\mathrm{Var}[\bar{Y}_n]$. For large n, the general formula for the variance of $\bar{Y}_n$ is [68]

$$\mathrm{Var}[\bar{Y}_n] \sim \nu(d) f(n^{-1}) n^{-1} \sim \nu(d) c_f n^{2d-1} \ , \ n \to \infty \ , \tag{2.5}$$

where

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

As $n \to \infty$, $f(n^{-1})$ corresponds with the spectral density at low frequencies, which takes the form (cf., Eq. 2.2)

$$f(\lambda) \sim \frac{\sigma^2}{2\pi} \frac{1 + \sum_{j=1}^{q} \theta_j}{1 - \sum_{i=1}^{p} \phi_i}^2 \cdot |\lambda|^{-2d} \sim c_f \cdot |\lambda|^{-2d}, \ \lambda \to 0 \ , \tag{2.6}$$

so that c_f is the pre-factor 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. (2.5), we compute the empirical variance-time relations [35] between various averaging sizes n and $\mathrm{Var}[\bar{Y}_n]$ 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 non-overlapping blocks of length n 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 $d = 0.16$
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 2.3 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 -1 in the log-log scale. For short memory ($d = 0$), Eq. (2.5) indicates a variance-time curve of $2\pi c_f n^{-1}$ as $n \to \infty$. 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 of $2\pi c_f$, manifesting as a vertical shift above the white-noise curve in log-log scale.

According to Eq. (2.5), 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** [69]. 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. (2.5) 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 which 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.

2.3 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. (2.5).

2.3.1 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—e.g., 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 autocor-

relation function or the variance-time relation. The intensity of any possible long memory can be assessed using any of numerous heuristic methods, including the **rescaled range method** [32], the **KPSS statistic** [70], the **rescaled variance (V/S) method** [71], the **detrended fluctuation analysis (DFA)** [72], **Haar wavelets** [73], and the **adaptive fractal analysis** [74]. If heuristic methods are indicative of long memory, then a more complete characterization of the processes' dependence structure may be obtained through the more rigorous technique of **maximum likelihood estimation**.

2.3.2 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 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 [75].

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 due to numerous inversions of the $n \times n$ covariance matrix. Fortunately, numerous approximate likelihood methods have been developed, including both Frequentist [68] and Bayesian [76] approaches. Here, we employ the elegant and computationally efficient spectral domain approximation to the Gaussian likelihood proposed by [77]. The Whittle estimator is asymptotically equivalent in distribution to the exact maximum likelihood estimator for Gaussian data [78] and yields asymptotically consistent and normally distributed parameter estimates for non-Gaussian data [79]. 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 [80, 81].

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

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=0}^{N} Y_t \exp(it\lambda) \right|^2 \tag{2.7}$$

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

$$Q(\eta) = \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda;\eta)} d\lambda + \int_{-\pi}^{\pi} \log f(\lambda;\eta) d\lambda \tag{2.8}$$

where η is the vector of unknown parameters, which contains the long-memory parameter d as well as any autoregressive and moving average coefficients, q.v., Eq. (2.3), and $f(\lambda;\eta)$ is the spectral density function given the parameters in η. The Whittle estimator is defined as the value of η that minimizes $Q(\eta)$; for computation, the integrals in Eq. (2.8) are replaced by the corresponding sums over Fourier frequencies.

2.3.3 Model selection

Maximum-likelihood-based methods like the Whittle estimation require specification of the precise parametric form of the model, i.e., the value of p and q and whether or not $d = 0$, which can invite bias under model misspecification [78]. 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, i.e., 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, e.g., 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 [82]. We follow [83], who found that the Bayesian information criterion (BIC) of [84] 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 [85]. The BIC is defined as

$$\text{BIC} = -2 \cdot \log(\tilde{L}) + \log(n) \cdot r \tag{2.9}$$

where $\tilde{L}$ is the maximized value of the likelihood function, i.e., 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 $\log(n) \cdot r$ 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.

2.3.4 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 a FARIMA(p, d, q) model has been selected and fit to the anomalous data via the procedures in Secs. 2.32.3.2 and 2.3.3. 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. (2.5), replacing d and f with their estimates, $\hat{d}$ and $\hat{f}$; the latter is the spectral density in Eq. (2.6) 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

$$\bar{Y}_n \pm z_{1-\alpha/2} \quad \overline{\nu(\hat{d})\hat{f}(n^{-1})n^{-1}} \tag{2.10}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution [68]. The **one-sided** upper $(1-\alpha)\times 100\%$ confidence bound is

$$\bar{Y}_n + z_{1-\alpha} \quad \overline{\nu(\hat{d})\hat{f}(n^{-1})n^{-1}}, \tag{2.11}$$

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

$$\bar{Y}_n - z_{1-\alpha} \quad \overline{\nu(\hat{d})\hat{f}(n^{-1})n^{-1}}. \tag{2.12}$$

In practice, since we are estimating $\hat{f}$, the normal approximation may not be very accurate at small sample sizes due to 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. [68] uses 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 Sec. 2.4, this inflation is essentially negligible.

2.4 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-year record of observed daily average surface temperatures at Potsdam, Germany. Since an average over at least ten years of daily observations for most common state variables (thirty years for precipitation) is a classical climate definition by the World Meteorological Organization [86], we focus on mean states of the four 30-year periods 1896–1925, 1926–1955, 1956–1985, 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. 2.4. The salient tasks are

1. Remove the seasonal cycle and non-stationary trends.
2. Assess the potential for long-memory with heuristic methods, e.g., 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.
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-year 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 non-stationary 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 [55, 87]. Removing arbitrary trend from data that are actually stationary is no safer than treating

non-stationary data as stationary, therefore we detrend data only given compelling evidence that a trend exists.

To detect possible trends, we adopt the approach of [40]. For each segment, we first remove a linear trend, then choose a best FARIMA model by the methods in Sec. 2.32.3.3. 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 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 period of the Potsdam data. We return to the issue of arbitrary trend in the data and its relationship to estimation of long memory in Sec. 2.5.

To assess the potential for long memory in the data, we check the variance-time relation in Fig. 2.5 and the first-order DFA [72] in Fig. 2.6. Since $\mathrm{Var}[\bar{Y}_n]$ is expected to decay as n^{2d-1} (Sec. 2.32.2.3), the asymptotic slope of the variance-time relation admits a heuristic estimate of long-memory intensity $\hat{d}_{\mathrm{VT}}$. The resulting estimates of the long-memory parameter, $\hat{d}_{\mathrm{VT}}$ and $\hat{d}_{\mathrm{DFA}}$, are given in Table 2.1. 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 [80]. 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 ARMA(p, q) models and a pool of candidate 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-year period, we fit 81 candidate ARMA models and 81 candidate FARIMA models using the Whittle method (q.v., Sec. 2.32.3.2), and we select the

best model from each pool according to the BIC (q.v., Sec. 2.32.3.3). The parameter estimates for the selected ARMA and FARIMA models for each 30-year period are shown in Tables 2.2 and 2.3, respectively.

Figure 2.7 shows the periodogram for each 30-year 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 [88] (see Appendix B). We use the visual diagnostic at this step because the summative approaches may favor a model 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 ATS (average-transform-smooth) method described by [89]. 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 non-overlapping 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. [89] 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 2.8 shows the diagnostic spectral residual plots for both ARMA and FARIMA models over the four 30-year periods of the Potsdam temperature data. The residuals are grouped into three frequency bands corresponding to distinct regimes of atmospheric variability: two days to two weeks, two weeks to one year, and one year to thirty years. The high frequency band from two days to two weeks corresponds with mesoscale to synoptic variability, the middle band of two weeks to one year corresponds with the subseasonal to seasonal scale as defined in [90], and the low frequency band corresponds with interannual to multidecadal scales. The distribution of residuals in each frequency band are summarized by a box plot; 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 one year. However, for the estimation of variance at long time scales, aptness at low frequencies is critical. Figure 2.8a shows that, in the 1896–1925 period, while the ARMA model performs well, the FARIMA model overestimates variability on time scales longer than one year, resulting in a negative residual distribution in that band. Figure 2.8b-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. 2.6 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. (2.10) 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 2.4 and Fig. 2.9 show the time mean of each 30-year period along with the 95% confidence intervals computed from both the selected ARMA and FARIMA models. Although the width of 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, e.g., a comparison of the mean temperature states for periods 1956–1985 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. [87] and [55] 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.

Comparing the periods 1926–1955 and 1986–2015 leads to the same contradiction between ARMA- and FARIMA- based inferences. Only 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 uncertainty characterization have emerged for both observational datasets [91] and climate model simulations [92]. 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-year period.

2.5 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, 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 STL (seasonal decomposition of time series using Loess [93]), a filtering procedure for decomposing a time series into trend, seasonal, and remainder components. Figure 2.10 shows the STL decomposition of the Potsdam temperature data in the 30-year period from 1986–2015. 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 Sec. 2.4 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 one year. This implies that removal of the STL seasonal and trend components effectively removed long memory from the Potsdam data.

To further isolate the 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 Sec. 2.4. 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.

In Sec. 2.4, 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.

2.6 Discussions 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 [24], who derived the variance of a finite time average of a first-order continuous-time autoregressive process in terms of its autocorrelation function. [25] and [94] 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 [24,94]. 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 [95]. In the Appendix, 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. [25], [28], [95], [96], and [29] 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 [97, 98]. Hence, in this work we use confidence intervals rather than hypothesis tests—following the lead of [99], 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, [100] (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 [80]. In addition, we provide an explicit formula (Eq. 2.5) 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-year dataset of daily average surface temperatures from Potsdam, Germany. MK16 used the entire 123-year record to calibrate their model and variance estimate, resulting in identical widths for the confidence intervals of the four 30-year 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, 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-year 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: ±0.34°C (1896–1925), ±0.48°C (1926–1955), ±0.41°C (1956–1985), and ±0.71°C (1986–2015) (see Table 2.4) versus ±0.5°C for all periods in MK16.

2.7 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 (Sec. 2.2).

- 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 (Secs. 2.3 and 2.4). 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-year mean states for the surface temperature at Potsdam, Germany (Sec. 2.4). 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-year period (1986–2015) relative to any of the three preceding 30-year 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 depend critically on the dependence structure assumed for atmospheric variability. As evidence of long-memory in climate data accumulates [37–39, 41, 42] 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.

2.8 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. (2.5) generalizes that for a pure short-memory climate process given in Eq. (10) and (11) of [28]. Katz provides the variance for the sample mean $\bar{X}_n$ of an AR(p) process as

$$\mathrm{Var}[\bar{X}_n] \sim \frac{V(\vec{\phi})}{n} \quad (n \to \infty) \tag{2.13}$$

where

$$V(\vec{\phi}) = \frac{\sigma^2}{[1 - \sum_{k=1}^{p} \phi_k]^2}. \tag{2.14}$$

Noting that the AR(p) process is FARIMA$(p, 0, 0)$ and using Eq. (2.6) for $f(n^{-1})$ we have

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

which is consistent with Eq. 2.13 and 2.14.

2.9 Appendix B: Summative Model Selection Strategies

Several summative approaches for choosing between candidate ARMA and FARIMA models are available, including the difference of BIC (ΔBIC), goodness-of-fit testing, the likelihood-ratio test, and the simulation-based selection strategy of [88]. 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 $\Delta\mathrm{BIC} \equiv \mathrm{BIC}_{\mathrm{ARMA}} - \mathrm{BIC}_{\mathrm{FARIMA}}$ [101, 102]. 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 [101], evidence is weak for $0 < |\Delta\mathrm{BIC}| < 2$, positive for $2 < |\Delta\mathrm{BIC}| < 6$, strong for $6 < |\Delta\mathrm{BIC}| < 10$, and very strong for $|\Delta\mathrm{BIC}| > 10$.

Another option is to use the goodness of fit (GOF) test of [103]. 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 sim-

plification of the more complex model (FARIMA) against the alternative hypothesis that it is not [104]. A notable limitation of this method is that the simpler model must be nested in the more complex model [102]. That is, it must be possible to obtain the simpler model by fixing the values of certain parameters in the more complex model, i.e., 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, [88] 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 $\mathrm{lr}_{\mathrm{obs}} = l_f(\hat{\Theta}|\mathbf{x}) - l_g(\hat{\Xi}|\mathbf{x})$ 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 $\mathrm{lr}_{\mathrm{obs}}$ with the distributions of lr_f and lr_g [105], which are obtained analogously to $\mathrm{lr}_{\mathrm{obs}}$ 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 $\mathrm{lr}_{\mathrm{obs}}$ 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 (Sec. 2.4) to choose between the best ARMA and best FARIMA models for each of the 30-year periods in the Potsdam temperature data (Table 2.5). For the period 1896–1925 the difference in BIC (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 [106] 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 2.11 shows the empirical cumulative distribution functions of $\mathrm{lr}_{\mathrm{FARIMA}}$ and $\mathrm{lr}_{\mathrm{ARMA}}$ along with $\mathrm{lr}_{\mathrm{obs}}$ for each 30-year period. For the periods 1896–1925 and 1926–1955, the distributions of $\mathrm{lr}_{\mathrm{FARIMA}}$ and $\mathrm{lr}_{\mathrm{ARMA}}$ are well-separated with $\mathrm{lr}_{\mathrm{obs}}$ falling as

a typical value in the distribution of $\text{lr}_{\text{FARIMA}}$ and a rare value in lr_{ARMA}, indicating the FARIMA models to be superior. The outcome is essentially the same for the period 1926–1955, 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 1956–1985, $\Delta\text{BIC} = 3.9$, 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 and neither model clearly preferred over the other. Since in both periods the two models are non-nested, 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 1956–1985 and 1986–2015. Table 2.5 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-year periods of the Potsdam temperature data.

Table 2.1.
Heuristic estimates of the long-memory parameter of the average daily surface temperature record in each 30-year period where $\hat{d}_{\mathrm{VT}}$ and $\hat{d}_{\mathrm{DFA}}$ are respectively obtained from the variance-time relation and the detrended fluctuation analysis. Standard deviations of the estimates based on linear regression are given in parentheses.

Period	$\hat{d}_{\mathrm{VT}}$	$\hat{d}_{\mathrm{DFA}}$
1896–1925	0.06 (0.02)	−0.02 (0.03)
1926–1955	0.10 (0.02)	0.15 (0.01)
1956–1985	0.08 (0.02)	0.06 (0.02)
1986–2015	0.15 (0.02)	0.07 (0.03)

Table 2.2.
Parameter estimates for the ARMA model (see Eq. 2.3) minimizing the BIC in each 30-year period.

Period	p	q	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2
1896-1925	1	2	0.759			0.134	-0.084
1926-1955	3	0	0.900	-0.193	0.077		
1956-1985	2	2	1.451	-0.505		-0.496	-0.217
1986-2015	2	2	1.509	-0.543		-0.558	-0.226

Table 2.3.
Parameter estimates for the FARIMA (see Eqs. 2.3 and 2.4) model minimizing the BIC in each 30-year 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
1926-1955	3	0	0.143	0.029	0.753	-0.145	0.055		
1956-1985	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		

Table 2.4.
Mean daily surface temperature at Potsdam for each of four 30-year periods ($\bar{Y}_{30\text{yr}}$) (°C) along with the variance of the mean (Var$[\bar{Y}_{30\text{yr}}]$), the half-width of the 95% confidence interval ($z_{0.975} \cdot \text{Var}[\bar{Y}_{30\text{yr}}]^{1/2}$), see Eq. (2.10), and lower/upper bounds of the 95% confidence interval (CI) estimated from the selected ARMA and FARIMA models.

Period	Model	$\bar{Y}_{30\text{yr}}$	Var$[\bar{Y}_{30\text{yr}}]$	95% CI half-width	95% CI lower	95% CI 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
1926-1955	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
1956-1985	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

Table 2.5.
Outcomes of methods for choosing between the ARMA and FARIMA model for each 30-year period. In columns two through six, results supporting the FARIMA model are in bold, results supporting the ARMA model are in italic, and neutral results are in regular typeface.

		GOF p-value		LRT		Chosen
Period	ΔBIC	ARMA	FARIMA	p-value	Rust (2007)	Model
1896-1925	*–0.9*	0.20	0.18	**0.004**	**FARIMA**	FARIMA
1926-1955	**15.5**	0.94	0.97	**<0.001**	**FARIMA**	FARIMA
1956-1985	**3.9**	0.36	0.47		impartial	FARIMA
1986-2015	**3.3**	0.32	0.31		impartial	FARIMA

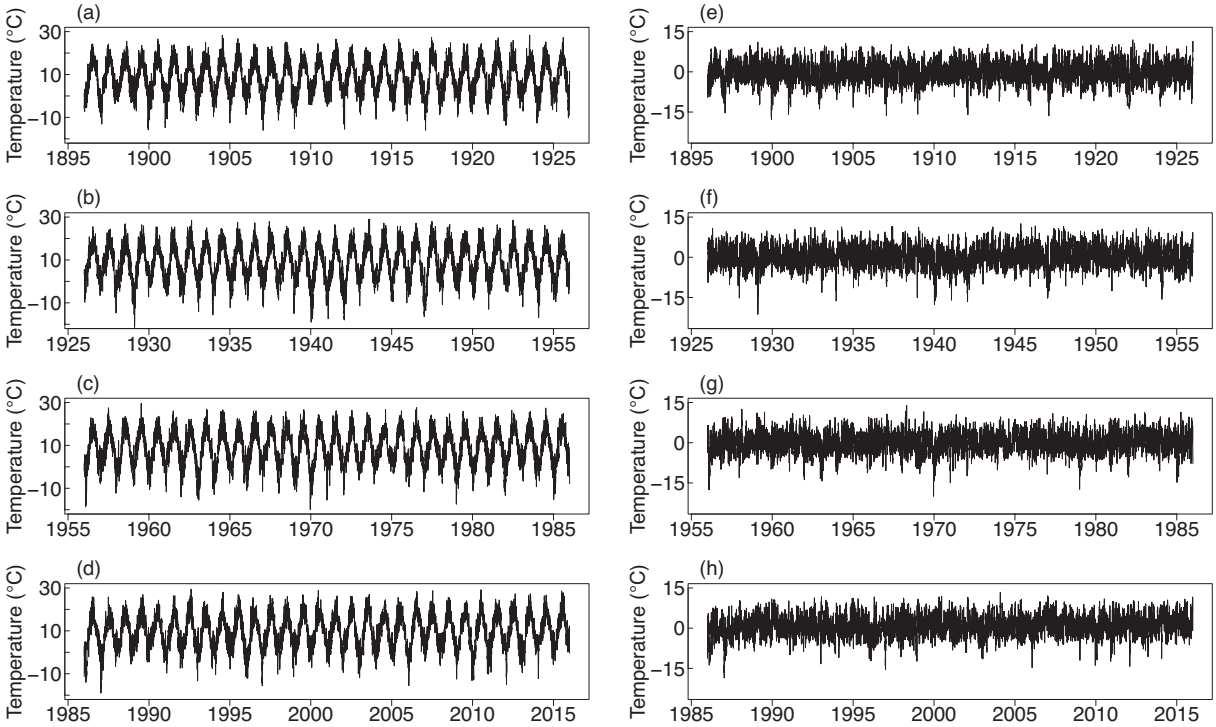

Figure 2.1. (a–d) raw Potsdam temperature data from the four 30-year periods: 1896–1925, 1926–1955, 1956–1985, and 1986–2015. (e–h) Potsdam temperature anomalies after removal of mean and seasonal cycle for the same periods.

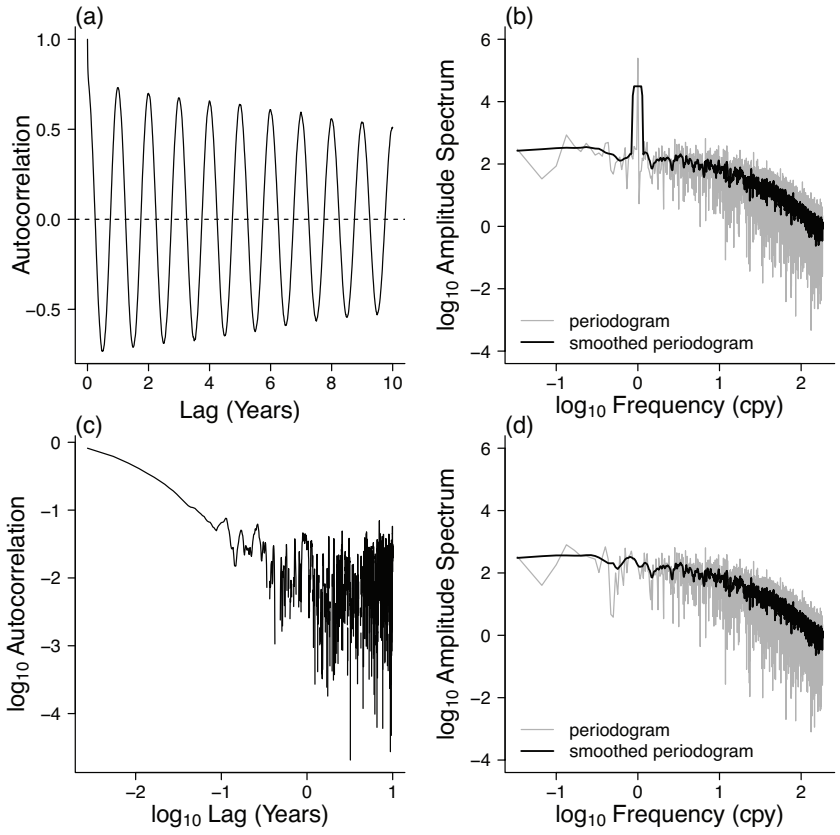

Figure 2.2. (a) Autocorrelation function and (b) periodogram of the raw Potsdam temperature data in the most recent period 1986–2015. (c) Autocorrelation function and (d) periodogram of the Potsdam temperature anomalies after removal of the seasonal cycle. Negative autocorrelation values (6% of the 3650 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.

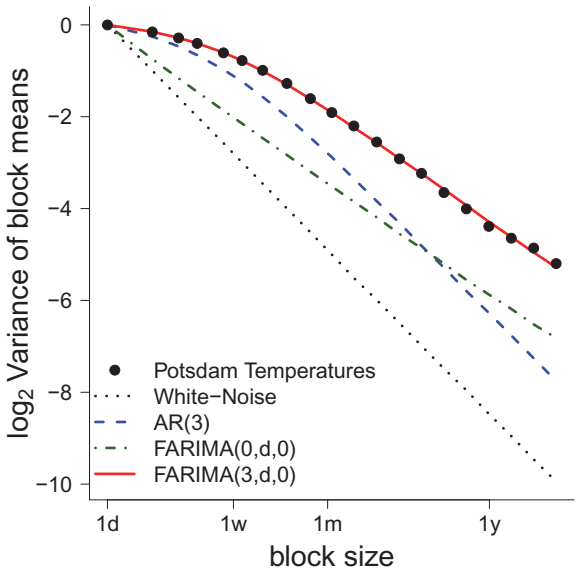

Figure 2.3. Empirical variance-time relations between sample variance of block-means and block size for the Potsdam temperature anomalies, independent white-noise, short-memory AR(3), long-memory FARIMA$(0, d, 0)$, and both short- and long-memory FARIMA$(3, d, 0)$.

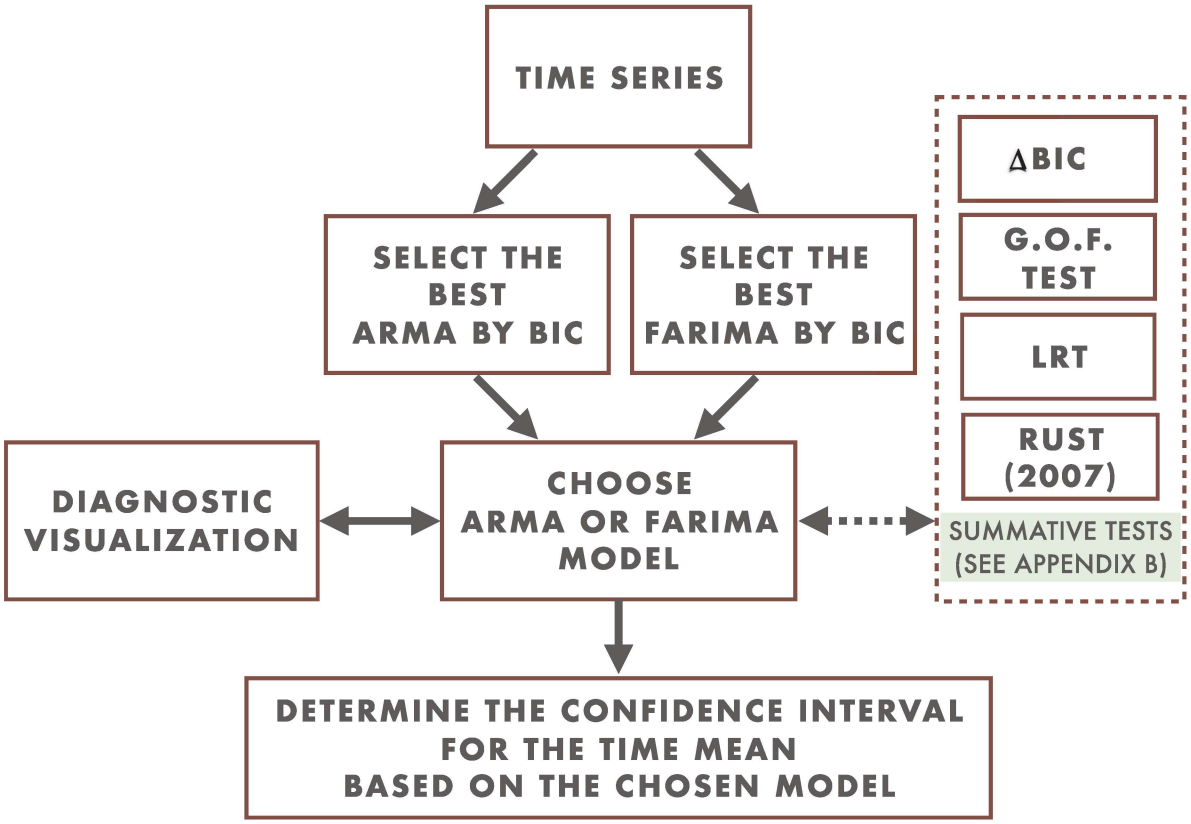

Figure 2.4. Schematic diagram illustrating the procedure for obtaining appropriate error bars for the mean of climate time series data. Enumeration corresponds to the tasks numbered in the text.

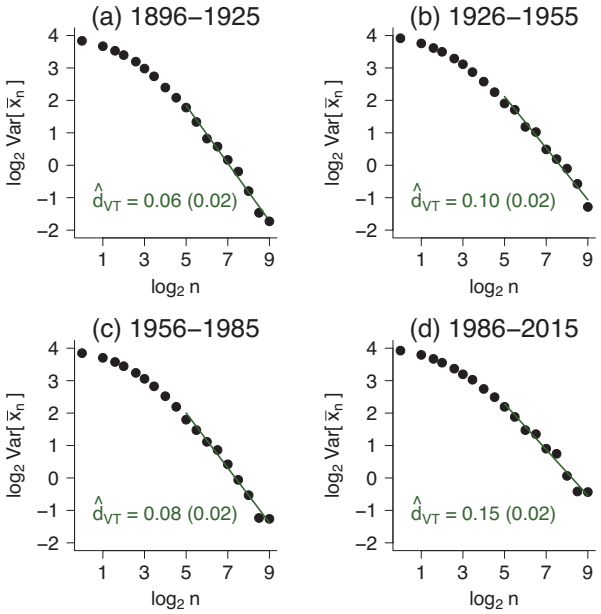

Figure 2.5. Variance-time plot of the Potsdam temperature anomalies within each of the 30-year periods on record. The regression line is fitted to scales $2^5 \leq m \leq 2^{10}$ (32–1024 days). The estimates $\hat{d}_{\mathrm{VT}}$ are based on the slope which corresponds to $2d - 1$, and the standard deviations of the estimates are given in parentheses.

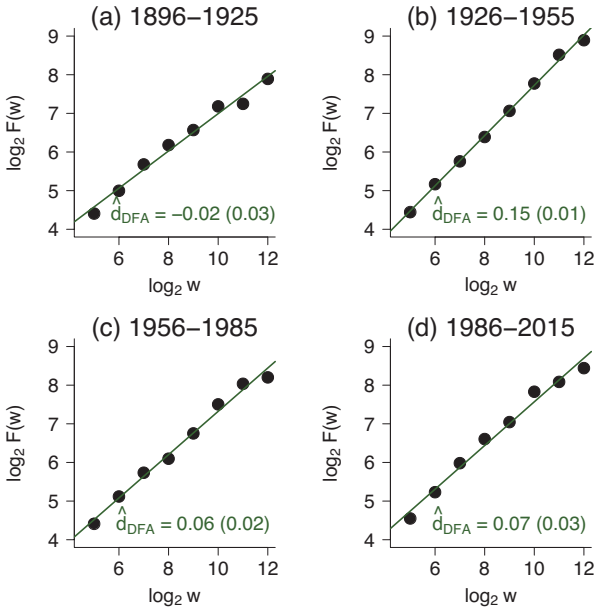

Figure 2.6. First-order detrended fluctuation analysis (DFA-1) of the Potsdam temperature anomalies within each of the 30-year periods on record. The regression line is fit to scales $w \geq 2^5$ (32 days) yielding the estimates $\hat{d}_{\mathrm{DFA}}$ and their standard deviations (in parentheses).

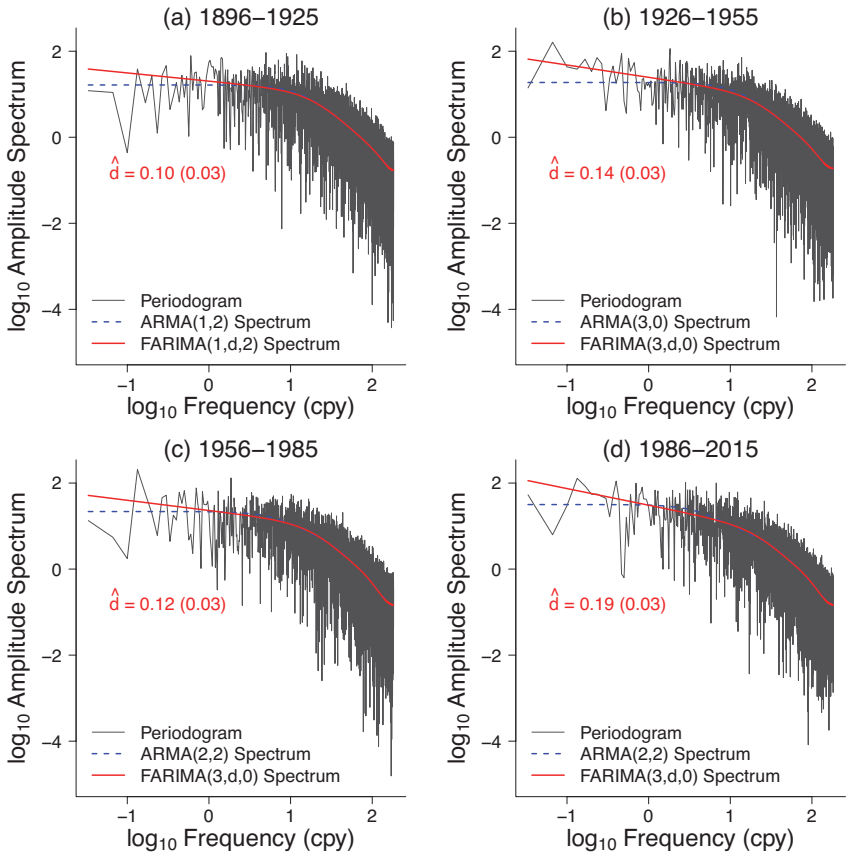

Figure 2.7. Periodogram of the Potsdam temperature anomalies along with the spectra of the selected FARIMA and ARMA models for each 30-year period on record. Estimates of the long-memory parameter d included in the FARIMA models are given along with their standard deviations in parentheses.

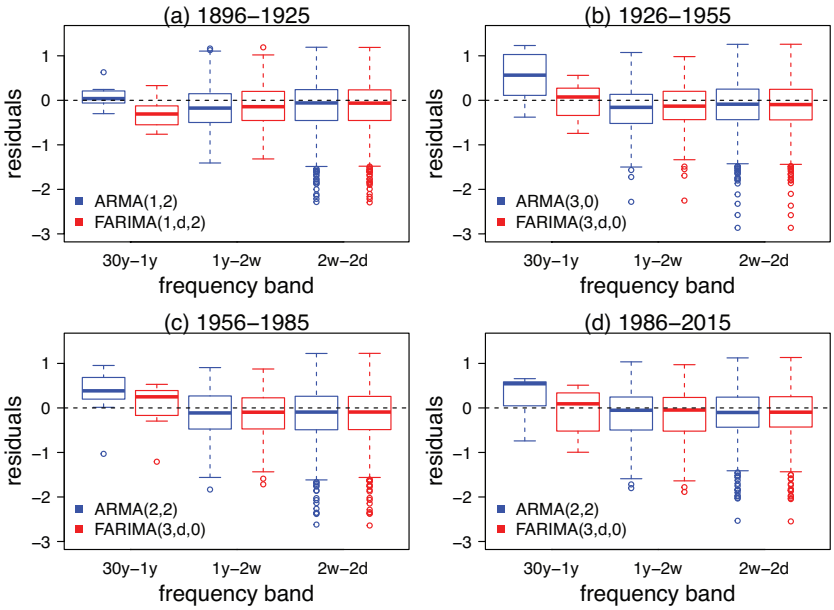

Figure 2.8. ATS-based spectral diagnostic visualization for the models fitted to the four 30-year periods of Potsdam temperature anomalies. The plots show the distribution of spectral residuals $\ln[\hat{I}(\lambda)] - \ln[\bar{f}(\lambda)]$ corresponding to both the selected ARMA and FARIMA models for each period.

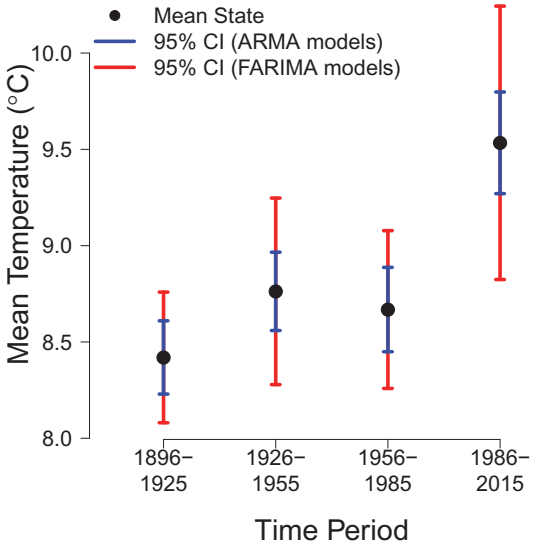

Figure 2.9. Mean states of the Potsdam temperature along with the 95% confidence intervals obtained assuming short-memory (ARMA model) and accounting for long-memory (FARIMA model) for each 30-year period on record.

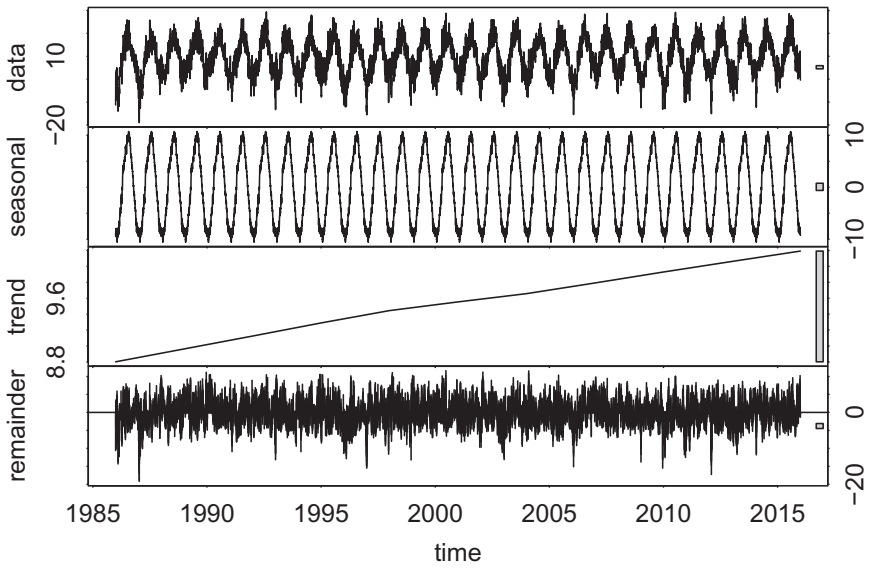

Figure 2.10. STL decomposition of the Potsdam temperature data in the period 1986–2015 into trend, seasonal, and remainder components. Vertical scales are in units of °C.

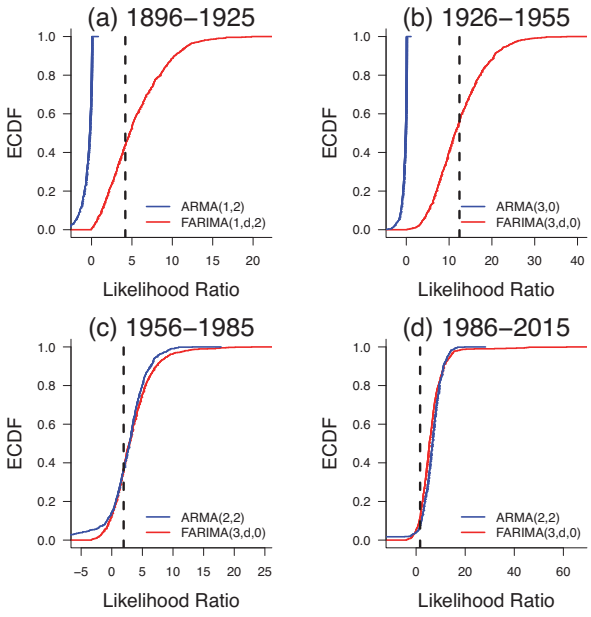

Figure 2.11. Empirical cumulative distribution functions (ECDF) of simulated likelihood ratios $\mathrm{lr}_{\mathrm{FARIMA}}$ and $\mathrm{lr}_{\mathrm{ARMA}}$ along with $\mathrm{lr}_{\mathrm{obs}}$ (vertical dashed line) for each 30-year period in the Potsdam temperature record.

3 RECENT CHANGES IN THE TEMPORAL CLUSTERING PATTERNS OF TROPICAL RAINFALL INFERRED FROM TRMM DATA

3.1 Introduction

Because precipitation controls the availability of water, changes in its established patterns have profound impacts on societies, economies, and ecosystems [107]. The increasing population living in water stressed conditions underscores the urgency for adaptation strategies based on scientific understanding of the changes in rainfall patterns [108, 109].

Changes in the hydroclimates of tropical and extratropical regions around the globe have been observed over the past few decades. In particular, reduced rainfall and increased frequency of drought have been observed in Australia [110], the Mediterranean region [111], the southwestern United States [112], the South American Altiplano [113], northern China [114], and in Africa and southeast Asia [115,116]. Rainfall and drought exhibit natural variability on multiple space and time scales, driven by teleconnections with interannual modes of climate variability such as the El Niño Southern Oscillation (ENSO) [117]. The recent trends in rainfall and drought may also be driven in part by changes in atmospheric circulation due to anthropogenic climate change [115].

Observations suggest that the global tropical belt has expanded by several degrees since the late 1970s [118–121] and that this expansion is expected to contribute to increased frequency of drought in the extratropics [122]. A number of studies have investigated the roles of various possible mechanisms including increased subtropical static stability [123], increased moisture transport by the mean circulation [124], direct radiative effects due to greenhouse gases and stratospheric ozone depletion [125], poleward shifts in extratropical jets [126], changes in extratropical circulation

[127], atmospheric heating due to black carbon and tropospheric ozone [128], and teleconnection with the Pacific Decadal Oscillation (PDO) [129, 130].

A particularly challenging aspect of tropical expansion research is the rather nebulous concept of the tropical edge, which resists precise definition due to its variability with season and geographic location. This nebulosity has lead to a number of edge definitions and thus a wide variety of estimated expansion rates [121, 131]. Discrepancies among estimated expansion rates have also been attributed to differences in reanalysis data products [132, 133]. There is also a systematic difference between trends measured from observational and reanalysis data with those in general circulation models [134, 135]. These discrepancies among reanalysis and model products highlights the need for tropical expansion studies based on observational data and constitutes a major motivation for the present work.

Several previous studies have found evidence of tropical expansion in observational precipitation data [128, 132, 136]. These studies have all used monthly rainfall data products and have focused on changes in total or mean precipitation. While seasonal mean precipitation is critical for local water resource management, other more subtle changes that go beyond mean precipitation patterns, such as extreme events, are also important. [137] and [138] have found an increase in extreme tropical rainfall events over the past few decades, associated with an intensification of the tropical hydrological cycle.

In addition to the magnitude of rainfall captured by its seasonal aggregate and its extreme tail behavior, the temporal correlation structure of precipitation also plays an important role in water resource management. Precipitation exhibits **persistent temporal scaling** [30, 31, 139], i.e. power-law decay in its autocorelation function, often called **long-range dependence** or **long memory**. Persistent temporal scaling is connected with extreme events and has even been used as a criterion for defining them [140–142]. Indeed, persistent temporal scaling has been observed in association with extreme tropical convective events [35]. A further feature implied by persistence

is temporal clustering of extreme events, in which weather and climate extremes tend to occur in intermittent clusters [143].

The focus of this study is to illuminate recent changes in observed tropical rainfall patterns that are more subtle but no less important than shifts in seasonal mean or in the frequency of extreme events. In particular we use observational precipitation data to study changes since the late 1990s in the subseasonal temporal clustering of tropical rainfall events. We discuss the precipitation dataset, the analytic method used to measure clustering, and the computational/statistical paradigm that drive the analysis in Sec. 3.2. We describe the results in Sec. 3.3, and we discuss the results and conclude in Sec. 3.4.

3.2 Data and Methods

3.2.1 TRMM Dataset

The Tropical Rainfall Measuring Mission (TRMM) Version 7 3B42 Multisatellite Precipitation Analysis (TMPA) data product combines calibrated precipitation estimates from multiple satellites and surface rain gauges where feasible [144]. The dataset has global spatial coverage from 50° S–50° N with $0.25° \times 0.25°$ horizontal resolution, spanning the period from 1998 through early 2015 with 3-hourly resolution. At fine time scales, the TMPA successfully reproduces the surface observation-based distribution of precipitation as well as large daily events. However, in common with other fine-scale rainfall estimators, it has lower skill in correctly specifying low and moderate rainfall amounts on short time scales [144]. We remove data before October 1998 and poleward of 40° latitudes because of a nontrivial fraction of missing data. The small fraction of remaining missing data points are imputed by linear interpolation.

3.2.2 Detrended fluctuation analysis

The persistent temporal scaling properties of a time series can be conveniently characterized by the Hurst exponent or Hurst parameter H [32]. The Hurst parameter quantifies the persistence of correlations such that when $0 < H < 1/2$, the signal has anti-persistent correlations; when $H = 1/2$, the signal is memoryless or has short-range correlations; when $1/2 < H < 1$, the signal has persistent long-range correlations; and when $H > 1$ the signal may be non-stationary or have non-trivial trends. While the Hurst parameter can be estimated by a variety of methods, few are capable of accurately estimating H when $H > 1$ [52, 145]. One of the few methods capable of accurate estimation in the presence of trends or non-stationarity is detrended fluctuation analysis (DFA) [72]. Because of the potential for non-stationarity associated with tropical expansion, we use DFA to estimate the strength of temporal scaling in the tropical rainfall data.

DFA works as follows: given a noise (or increment) time series, $x_1, x_2, x_3, \cdots$, with mean $\overline{x}$, one first constructs a random walk process,

$$u(i) = \sum_{k=1}^{i} (x_k - \overline{x}), \quad i = 1, 2, \cdots, N \tag{3.1}$$

then divides $\{u(i),\ i = 1, 2, \cdots, N\}$ into N/m non-overlapping segments (where N/m denotes the largest integer equal to or smaller than N/m), each containing m points. The local trend in each segment is then computed, typically as the ordinate of a best linear or polynomial fit of the random walk in that segment. Finally, one computes the "detrended walk", denoted by $u_m(k), k = 1, 2, ..., m$, as the difference between the original segment (the"walk"), $u(k)$, and the local trend. The fractal behavior is described by the following scaling law

$$F_d(m) = \left\langle \sum_{i=1}^{m} u_m(i)^2 \right\rangle^{1/2} \sim m^H \tag{3.2}$$

where the angle brackets denote ensemble averages of all the segments.

While it is common to use linear or polynomial regression to estimate the local trend in each segment, this leads to discontinuities or even large abrupt jumps at the

segment boundaries. By using a smooth trend instead of the discontinuous piecewise trend, DFA can better handle non-stationarity or arbitrary trends in the data [74]. Here, we use the Loess smoother [146] to compute a globally smooth trend $v_w(i)$, where w denotes the span (bandwidth) of the Loess smoother. The residual, $u(i) - v_w(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to

$$F(w) = \frac{1}{N} \sum_{i=1}^{N} (u(i) - v_w(i))^2 \Big)^{1/2} \sim w^H. \tag{3.3}$$

To capture the clustering of rainfall events within each 6-month season, we estimate the fluctuation function $F(w)$ on the time scales from 12 hours to 16 days, which correspond with mesoscale to synoptic scale atmospheric variability. To estimate H, we use the slope coefficient of a linear regression of $\log_2 F(w)$ against $\log_2 w$ [**?**, e.g.,]]bowers2013long.

3.2.3 Divide and Recombine

To scrutinize the temporal correlation structure in the TRMM precipitation data, we employ the analysis framework and computational tools of the DeltaRho Project (http://www.deltarho.org). The framework, called *divide and recombine*, involves dividing a large complex dataset into subsets and applying analytic or visualization methods to the subsets before statistical, analytical, or graphical recombination of the results [147, 148].

Our primary interest is in the temporal correlation structure of local precipitation. Since precipitation regimes can vary dramatically across locations and seasons, we divide the dataset into subsets, each consisting of a 6-month-long 3-hourly time series of precipitation rates for a particular location. We use the concept of "monsoon years [149]" starting with summer as May through October, followed by winter as November through the next April. For each of the 462240 locations, we have 17 winter subsets (winter 1998 to winter 2014) and 16 summer subsets (summer 1999 to summer 2014), $\sim$ 15 million subsets in total. We apply DFA to the time series

data in each subset to estimate the persistence (as discussed in Sec. 3.3) yielding an estimate of H for each subset. We perform several statistical recombinations of the DFA results. After applying DFA, we perform a statistical recombination of the results for each location and season, e.g., taking an average, yielding a map of that statistic for both summer and winter seasons. This division, analysis, and statistical recombination are carried out in massive parallel on a Hadoop cluster running the DeltaRho software stack.

3.3 Results and Discussions

Figure 3.1 shows two example seasonal subset precipitation time series. Both subsets have a seasonal average precipitation rate of about 3 mm day^{-}1; however, the distribution of rainfall throughout the season is markedly different. Whereas the top panel shows a somewhat homogenous distribution of rainfall throughout the season, the bottom panel shows a strong clustering of rainfall into a relatively short period of time. While the two subsets have the same seasonal average water availability, the subset in the bottom panel may be considered to experience drought, while that in the top panel does not.

The difference in the temporal distribution of rainfall between the two subsets in Fig. 3.1 is echoed by the estimates of Hurst parameter. The top panel exhibits a homogenous temporal distribution or weak clustering and has H near 1/2, indicating a memoryless precipitation process. On the other hand, the bottom panel which exhibits strong clustering has $H \approx 0.9$ which indicates persistence in the precipitation process. The persistent clustering exists on multiple scales; on a seasonal scale, the majority of rainfall is confined to the month from mid November to mid December, but on a weekly scale the rainfall within that month is also clustered into distinct events. This fractal behavior is consistent with other findings of self-similarity in rainfall [**?**, e.g.,]]lovejoy1985fractal. We note that while these two locations have equivalent seasonal average rainfall, the difference in temporal clustering patterns

echoed by the Hurst parameter has profoundly different implications for local water resource management.

Having established that the Hurst parameter can distinguish various rainfall clustering patterns we now examine the average spatial distribution of clustering patterns in the tropical belt. Figure 3.2 shows the average Hurst parameter for each location and season. Each pixel in the figure represents the average of the 17 winter or 16 summer H values estimated at each location. In general, H values range between 1/2 and 1 indicating that tropical precipitation is either memoryless or persistent on meso and synoptic time scales. There is a pronounced land-sea contrast with a tendency for marine rainfall to be persistent and land rainfall to be memoryless. Persistence tends to be enhanced in convectively active regions such as the Indian Pacific warm-pool, the Intertropical Convergence Zone, the South Pacific Convergence Zone, the tropical Atlantic, and the Caribbean. Persistence also tends to be particularly strong in regions with sharp horizontal gradients in mean precipitation, as indicated by the 2 mm/day and 4 mm/day contours.

It is possible that the scaling behavior measured on scales from 12 hours to 16 days is better defined in some locations than others. In particular, the scaling exponent H can actually take different values over different frequency bands; the transition point between scaling regimes is called a **scaling break** [74] or **crossover** [139]. To obtain a heuristic indication of possible scaling breaks we use the coefficient of determination R^2 from the linear models used to estimate H. Low R^2 indicates that there is variability in $\log_2 F(w)$ that is not accounted for by the single linear fit, i.e., a scaling break.

Figure 3.3 shows the seasonal average maps of R^2. The seasonal average R^2 are obtained analogously to the mean H in Fig. 3.2, i.e., they are the average R^2 of either 17 winter subsets or 16 summer subsets for each location. In general, the fit quality indicated by R^2 tends to be very high in regions with high seasonal average precipitation as well as over land, especially in the summer hemisphere. There is some tendency for regions with low mean precipitation to have lower quality fits.

This could be indicative of the presence of scaling breaks in the band from 12 hours to 16 days. It could also reflect increased sampling variability in the fluctuation function $F(w)$ due to the smaller number of positive precipitation rate observations available in certain arid locations. Further diagnosis of the additional scaling regimes that could be present in locations with lower quality fits is an excellent subject for future investigations.

In order to examine changes in the persistence and temporal clustering patterns of tropical rainfall, we measure the intensity of trends in the seasonal Hurst parameter at each location. This constitutes a statistical recombination in which the Hurst parameters estimated in the 17 winter subsets or 16 summer subsets are linearly regressed against time, yielding a slope estimate giving the average annual change in H for each season at each location. Figure 3.4 shows locations with substantial change (more than 0.02 per year) in seasonal H over the period of record. For reference, regions with less than 0.04 mm/hr average precipitation are shaded.

We find locations with large changes in clustering patterns in arid coastal regions around the world, including the southwest US, the South American Altiplano, Africa, the Middle East, India, and southeast Asia. These regions have all been identified as water-stressed by either physical or economic water scarcity by the International Water Management Institute [150].

Such substantive changes of rainfall clustering in regions already experiencing water scarcity pose a major risk to inhabitants who's water management practices are designed for established rainfall patterns. Indeed [151] found that long-term average water availability is not strongly correlated with child malnutrition, whereas volatility in water availability is the single most important factor. The reason that long-term average water availability is less important is that the regions tend to be well adapted to the established precipitation patterns, employing agricultural practices such as transhumant and pastoralist systems. However, shocks in availability, such as drought or extreme rainfall, tend to stymie these adaptive practices and lead to depletion of household resources [152].

3.4 Conclusions

Our study shows that changes in observed tropical rainfall patterns in recent decades transcend changes in seasonal average precipitation. In particular, we illuminate the persistent temporal clustering patterns of rainfall through their correlation structure as quantified by the Hurst parameter H. These clustering patterns are of great practical importance for water resource management, since they control how the local seasonal water supply becomes available over time. We point out that while two regions may have very similar seasonal average precipitation, one may have very consistent rainfall while another has an envelope of extreme rainfall events followed by drought. We demonstrate that the Hurst parameter conveniently discriminates between these patterns and can therefore be used to characterize temporal clustering in rainfall.

We show that the seasonal average clustering patterns feature a land-sea contrast with much stronger clustering over the oceans. High temporal clustering tends to occur in regions with strong horizontal gradient of the mean precipitation and near coastal boundaries.

In light of the evidence that the tropical belt has expanded over recent decades, we also determine locations which have experienced large shifts in temporal clustering patterns from 1998–2014. While locations with both substantially increased and decreased clustering can be identified, the locations with greatest change in clustering predominantly fall in the water stressed regions of the world [150]. Even in cases where local mean precipitation is not shifting, the magnitude of the changes in clustering intensity constitute a substantial risk to local inhabitants whose water management practices are based on established patterns of rainfall. This underscores the need to go beyond basic summary statistics such as the mean when considering the potential impacts of widespread changes in the tropical hydroclimate.

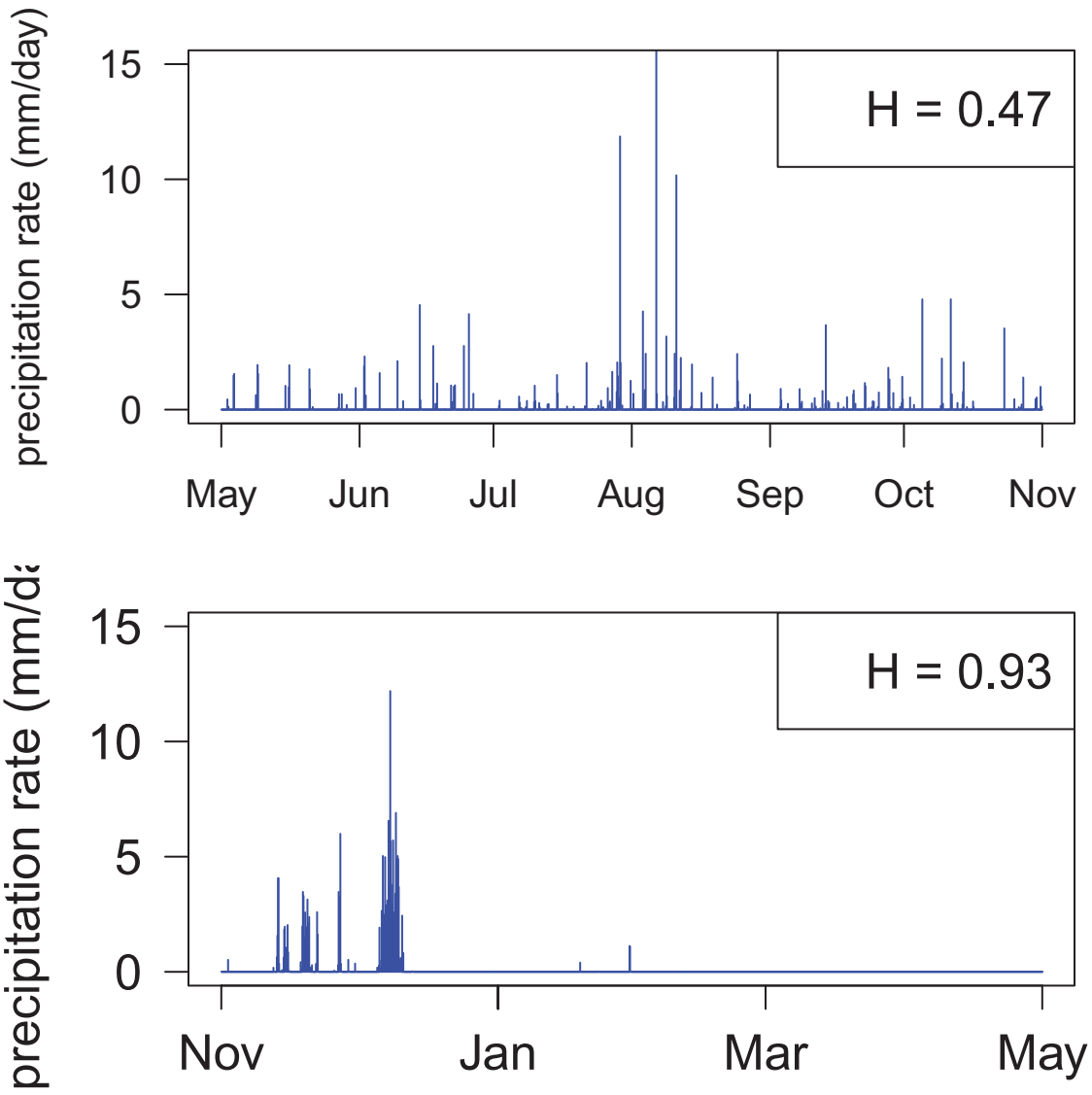

Figure 3.1. Seasonal block precipitation time series for (top) summer 2004 at 32.875°E, 7.875°N and (bottom) winter 2013 at 85.625°E, 14.625°N.

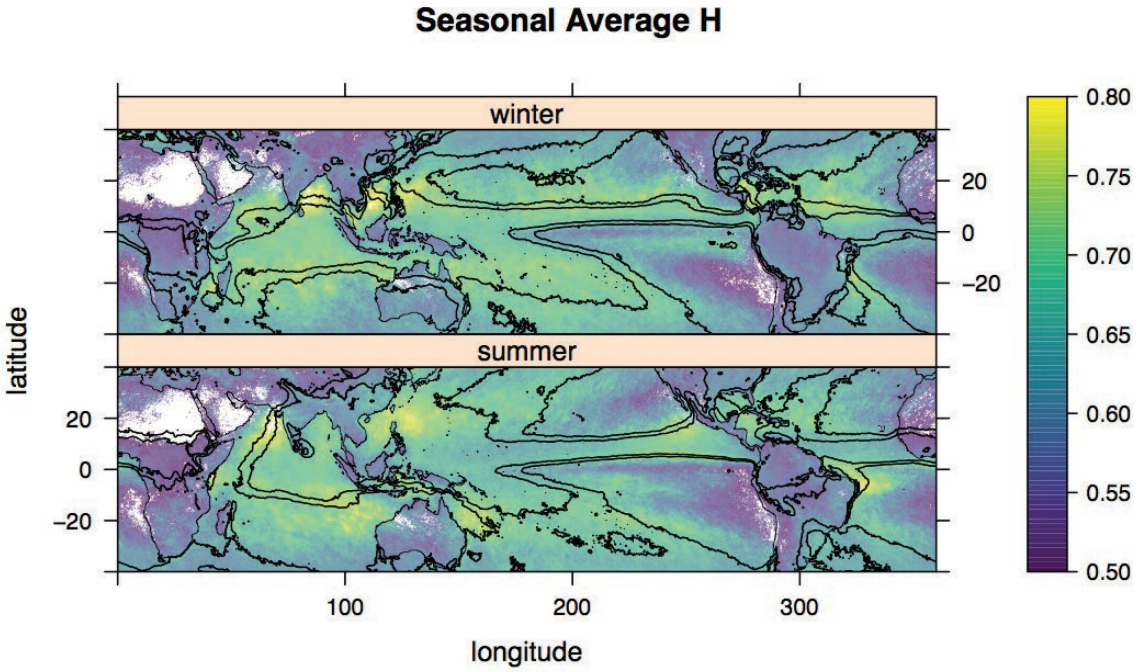

Figure 3.2. Map of seasonal average H. Contours indicate 2 and 4 mm/day average precipitation. White indicates locations where H is not defined.

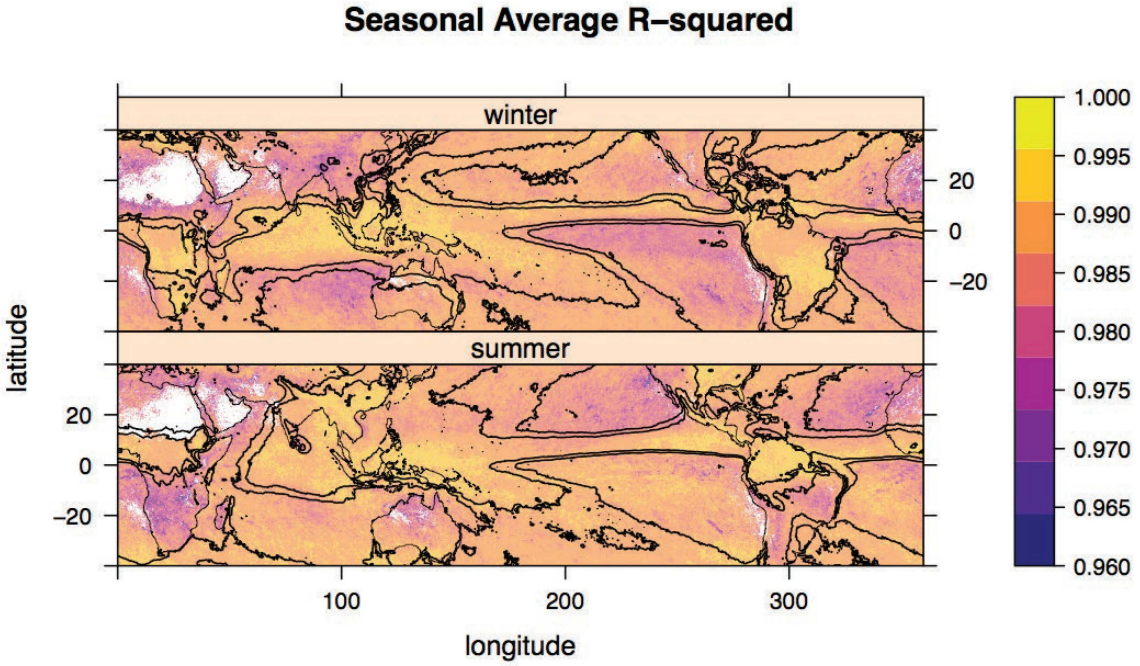

Figure 3.3. Map of seasonal average R^2 obtained in fitting H. Contours indicate 2 and 4 mm/day average precipitation. White indicates locations where H is not defined.

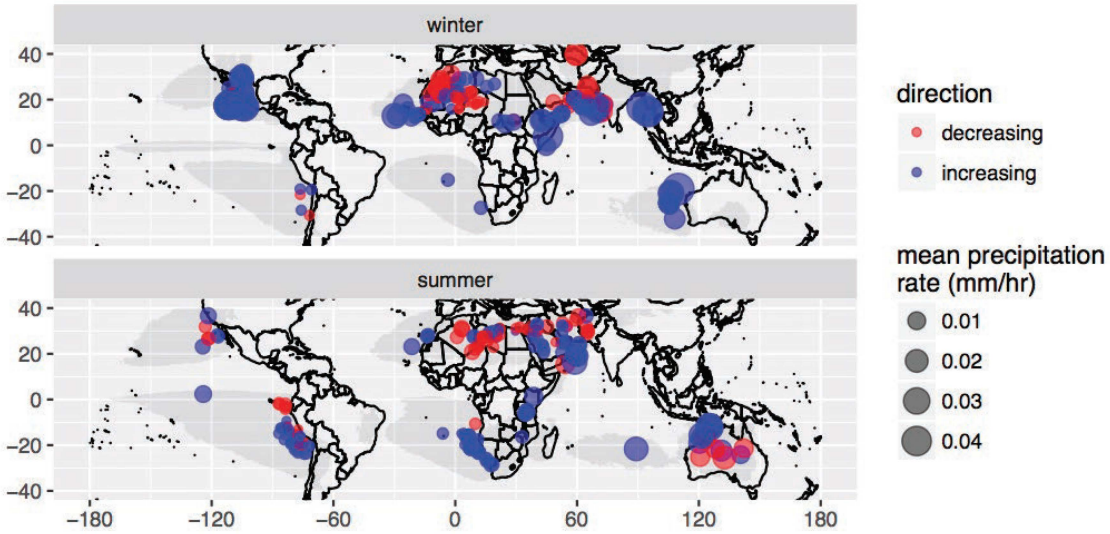

Figure 3.4. Locations with time change of H greater than 0.02 or less than -0.02 (/year). Regions with less than 0.04 mm/hr mean precipitation rate are shaded.

REFERENCES

REFERENCES

[1] A.M. García-Suárez, C.J. Butler, and M.G.L. Baillie. Climate signal in tree-ring chronologies in a temperate climate: A multi-species approach. *Dendrochronologia*, 27:183–198, 2009.

[2] T. Scharnweber, M. Manthey, C. Criegee, A. Bauwe, C. Schröder, and M. Wilmking. Drought matters–declining precipitation influences growth of *fagus sylvatica l.* and *quercus robur l.* in north-eastern Germany. *Forest Ecology and Management*, 262:947–961, 2011.

[3] M. E. Mann, Z. Zhang, M. K. Hughes, R. S. Bradley, S. K. Miller, S. Rutherford, and F. Ni. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences*, 105(36):13252–13257, 2008.

[4] H.F. Zhu, Y.H. Zheng, X.M. Shao, X.H. Liu, Y. Xu, and E.Y. Liang. Millennial temperature reconstruction based on tree-ring widths of Qilian juniper from Wulan, Qinghai Province, China. *Chinese Science Bulletin*, 53(24):3914–3920, 2008.

[5] Z. S. Li, Q. B. Zhang, and K. Ma. Tree-ring reconstruction of summer temperature for A.D. 14752003 in the central Hengduan Mountains, Northwestern Yunnan, China. *Climatic Change*, 110:455–467, 2011.

[6] C. L. Quesne. Ancient *austrocedrus* tree-ring chronologies used to reconstruct central Chile precipitation variability from A.D. 1200 to 2000. *Journal of Climate*, 19:5731–5744, 2006.

[7] C. B. Griggs, A. T. DeGaetano, P. I. Kuniholm, and M. W. Newton. A regional high-frequency reconstruction of May-June precipitation in the north Aegean from oak tree-rings, A.D. 1089-1989. *International Journal of Climatology*, 27(8):1075–1089, 2007.

[8] X. Shao, Y. Xu, z. Y. Yin, E. Liang, H. Zhu, and S. Wang. Climatic implications of a 3585-year tree-ring width chronology from the northeastern Qinghai-Tibetan Plateau. *Quaternary Science Reviews*, 29:2111–2122, 2010.

[9] C. Loehle. A mathematical analysis of the divergence problem in dendroclimatology. *Climatic Change*, 94:233–245, 2009.

[10] J. F. Eichner, E. Koscielny-Bunde, S. Havlin, and H. J. Schellnhuber. Power-law persistence and trends in the atmosphere: a detailed study of long temperature records. *Phys. Rev. E*, 68, 2003. 046133.

[11] S. Lovejoy and B. Mandelbrot. Fractal properties of rain and a fractal model. *Tellus*, 37A:209–232, 1985.

[12] G. Pandey, S. Lovejoy, and D. Schertzer. Multifractal analysis of daily river flows including extremes for basins of five to two million square kilometers, one day to 75 years. *J. Hydrol.*, 208:62–81, 1998.

[13] D Lavallée and H. Beltrami. Stochastic modeling of climatic variability in dendrochronology. *Geophys. Res. Lett.*, 31(L15202), 2004.

[14] L. Telesca and M. Lovallo. Long-range dependence in tree-ring width time series of *austrocedrus chilensis* revealed by means of the detrended fluctuation analysis. *Physica A*, 389:4096–4104, 2010.

[15] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineering*, 116:770–808, 1951.

[16] J.B. Gao, J. Hu, W.W. Tung, Y. Cao, N. Sarshar, and V.P. Roychowdhury. Assessment of long-range correlation in time series: How to avoid pitfalls. *Phys. Rev. E*, 73(016117), 2006.

[17] J. B. Gao, Y. H. Cao, W. W. Tung, and J. Hu. *Multiscale Analysis of Complex Time Series — Integration of Chaos and Random Fractal Theory, and Beyond*. Wiley, August 2007.

[18] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, and H.E. Stanley. Mosaic organization of DNA nucleotides. *Phys. Rev. E*, 49:1685–1689, 1994.

[19] J.B. Gao, J. Hu, and W.W. Tung. Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLoS ONE*, 6(9):e24331, 2011.

[20] E.R. Cook and R. L. Holmes. *Users Manual for Program ARSTAN, in Tree-Ring Chronologies of Western North America: California, eastern Oregon and northern Great Basin, by R. L. Holmes, R. K. Adams and H. C. Fritts*, pages 50–65. Laboratory of Tree-Ring Research, The University of Arizona, 1986.

[21] E. R. Cook, K. R. Briffa, D. M. Meko, D. A. Graybill, and G. Funkhouser. The ''segment length curse" in long tree-ring chronology development for paleoclimatic studies. *The Holocene*, 5:229–237, 1995.

[22] Keith R. Briffa and Thomas M. Melvin. A closer look at regional curve standardization of tree-ring records: Justification of the need, a warning of some pitfalls, and suggested improvements in its application. In Malcolm K. Hughes, Thomas W. Swetnam, and Henry F. Diaz, editors, *Dendroclimatology*, volume 11 of *Developments in Paleoenvironmental Research*, pages 113–145. Springer Netherlands, 2011.

[23] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, third edition, 2002.

[24] C E Leith. The standard error of time-average estimates of climatic means. *Journal of Applied Meteorology*, 12(6):1066–1069, 1973.

[25] Richard H Jones. Estimating the variance of time averages. *Journal of Applied Meteorology*, 14(2):159–163, 1975.

[26] R Madden and W Sadeh. Empirical estimates of the standard error of time-averaged climatic means. *Journal of Applied Meteorology*, 14(2):164–169, 1975.

[27] Roland A Madden. Estimates of the natural variability of time-averaged sea-level pressure. *Monthly Weather Review*, 104(7):942–952, 1976.

[28] Richard W Katz. Statistical evaluation of climate experiments with general circulation models: A parametric time series modeling approach. *Journal of the Atmospheric Sciences*, 39(7):1446–1455, 1982.

[29] Francis W Zwiers and Hans von Storch. Taking serial correlation into account in tests of the mean. *Journal of Climate*, 8(2):336–351, 1995.

[30] S Lovejoy and B B Mandelbrot. Fractal properties of rain, and a fractal model. *Tellus A*, 37(3):209–232, 1985.

[31] Jan W Kantelhardt, Eva Koscielny-Bunde, Diego Rybski, Peter Braun, Armin Bunde, and Shlomo Havlin. Long-term persistence and multifractality of precipitation and river runoff records. *Journal of Geophysical Research: Atmospheres*, 111(D1), 2006.

[32] Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.*, 116:770–808, 1951.

[33] Jon D Pelletier and Donald L Turcotte. Long-range persistence in climatological and hydrological time series: analysis, modeling and application to drought hazard assessment. *Journal of Hydrology*, 203(1):198–208, 1997.

[34] G Pandey, S Lovejoy, and D Schertzer. Multifractal analysis including extremes of daily river flow series for basins one to a million square kilomemeters. *Journal of Hydrology*, 208:62–81, 1998.

[35] Wen-wen Tung, Mitchell W. Moncrieff, and Jian-Bo Gao. A systemic analysis of multiscale deep convective variability over the tropical pacific. *Journal of Climate*, 17(14):2736–2751, 2004.

[36] A A Tsonis, P J Roebber, and J B Elsner. Long-range correlations in the extratropical atmospheric circulation: Origins and implications. *Journal of Climate*, 12(5):1534–1541, 1999.

[37] Dmitry I Vyushin and Paul J Kushner. Power-law and long-memory characteristics of the atmospheric general circulation. *Journal of Climate*, 22(11):2890–2904, 2009.

[38] Eva Koscielny-Bunde, Armin Bunde, Shlomo Havlin, H Eduardo Roman, Yair Goldreich, and Hans-Joachim Schellnhuber. Indication of a universal persistence law governing atmospheric variability. *Physical Review Letters*, 81(3):729, 1998.

[39] Peter Huybers and William Curry. Links between annual, milankovitch and continuum temperature variability. *Nature*, 441(7091):329–332, 2006.

[40] Christian Franzke. Long-range dependence and climate noise characteristics of antarctic temperature data. *Journal of Climate*, 23(22):6074–6081, 2010.

[41] Christian Franzke. Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of Climate*, 25(12):4172–4183, 2012.

[42] Naiming Yuan, Minghu Ding, Yan Huang, Zuntao Fu, Elena Xoplaki, and Juerg Luterbacher. On the long-term climate memory in the surface air temperature records over antarctica: A nonnegligible factor for trend evaluation. *Journal of Climate*, 28(15):5922–5934, 2015.

[43] ER Thomas, PF Dennis, Thomas J Bracegirdle, and Christian Franzke. Ice core evidence for significant 100-year regional warming on the antarctic peninsula. *Geophysical Research Letters*, 36(20), 2009.

[44] M C Bowers, Jianbo Gao, and Wen-wen Tung. Long range correlations in tree ring chronologies of the usa: Variation within and across species. *Geophysical Research Letters*, 40(3):568–572, 2013.

[45] William H Klein. A hemispheric study of daily pressure variability at sea level and aloft. *Journal of Meteorology*, 8(5):332–346, 1951.

[46] Roland A Madden and Dennis J Shea. Estimates of the natural variability of time-averaged temperatures over the united states. *Monthly Weather Review*, 106(12):1695–1703, 1978.

[47] D S Gutzler and K C Mo. Autocorrelation of northern hemisphere geopotential heights. *Monthly Weather Review*, 111(1):155–164, 1983.

[48] Rudolf O Weber and Peter Talkner. Spectra and correlations of climate data from days to decades. *Journal of Geophysical Research: Atmospheres*, 106(D17):20131–20144, 2001.

[49] Rodrigo Caballero, Stephen Jewson, and Anders Brix. Long memory in surface air temperature detection, modeling, and application to weather derivative valuation. *Climate Research*, 21(2):127–140, 2002.

[50] Jan F Eichner, Eva Koscielny-Bunde, Armin Bunde, Shlomo Havlin, and H-J Schellnhuber. Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. *Physical Review E*, 68(4):046133, 2003.

[51] Luis A Gil-Alana. Statistical modeling of the temperatures in the northern hemisphere using fractional integration techniques. *Journal of Climate*, 18(24):5357–5369, 2005.

[52] Jianbo Gao, Jing Hu, Wen-Wen Tung, Yinhe Cao, N Sarshar, and Vwani P Roychowdhury. Assessment of long-range correlation in time series: how to avoid pitfalls. *Physical Review E*, 73(1):016117, 2006.

[53] S Lennartz and A Bunde. Trend evaluation in records with long-term memory: Application to global warming. *Geophysical Research Letters*, 36(16), 2009.

[54] Sabine Lennartz and Armin Bunde. Distribution of natural trends in long-term correlated records: A scaling approach. *Physical Review E*, 84(2):021129, 2011.

[55] Josef Ludescher, Armin Bunde, Christian LE Franzke, and Hans Joachim Schellnhuber. Long-term persistence enhances uncertainty about anthropogenic warming of antarctica. *Climate dynamics*, 46(1-2):263–271, 2016.

[56] George E P Box and Gwilym M and Jenkins. *Time series analysis: Forecastingand control*. Holden Day, San Francisco, 1970.

[57] Richard W Katz and Richard H Skaggs. On the use of autoregressive-moving average processes to model meteorological time series. *Monthly Weather Review*, 109(3):479–484, 1981.

[58] Ian T Jolliffe. Quasi-periodic meteorological series and second-order autoregressive processes. *Journal of Climatology*, 3(4):413–417, 1983.

[59] Tiao J Chang, M L Kavvas, and J W Delleur. Daily precipitation modeling by discrete autoregressive moving average processes. *Water Resources Research*, 20(5):565–580, 1984.

[60] Johnny Wei-Bing Lin and J David Neelin. Considerations for stochastic convective parameterization. *Journal of the Atmospheric Sciences*, 59(5):959–975, 2002.

[61] Ergin Erdem and Jing Shi. Arma based approaches for forecasting the tuple of wind speed and direction. *Applied Energy*, 88(4):1405–1414, 2011.

[62] Pierre Ailliot and Valérie Monbet. Markov-switching autoregressive models for wind time series. *Environmental Modelling & Software*, 30:92–101, 2012.

[63] Clive WJ Granger and Roselyne Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980.

[64] Jonathan R M Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 1981.

[65] Alberto Montanari, Renzo Rosso, and Murad S Taqqu. Some long-run properties of rainfall records in italy. *Journal of Geophysical Research: Atmospheres*, 101(D23):29431–29438, 1996.

[66] J-S von Storch, P Müller, and E Bauer. Climate variability in millennium integrations with coupled atmosphere-ocean gcms: A spectral view. *Climate Dynamics*, 17(5-6):375–389, 2001.

[67] Rajesh G Kavasseri and Krithika Seetharaman. Day-ahead wind speed forecasting using f-arima models. *Renewable Energy*, 34(5):1388–1393, 2009.

[68] Jan Beran, Yuanhua Feng, Sucharita Ghosh, and Rafal Kulik. *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Springer Berlin Heidelberg, 2013.

[69] James Wallis and B Mandelbrot. Noah, joseph and operational hydrology. *Water Resources Research*, 4(5):909–918, 1968.

[70] Denis Kwiatkowski, Peter C B Phillips, Peter Schmidt, and Yongcheol Shin. 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? *Journal of Econometrics*, 54(1):159–178, 1992.

[71] Liudas Giraitis, Piotr Kokoszka, Remigijus Leipus, and Gilles Teyssiere. Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics*, 112(2):265–294, 2003.

[72] C-K Peng, Sergey V Buldyrev, Shlomo Havlin, Michael Simons, H Eugene Stanley, and Ary L Goldberger. Mosaic organization of dna nucleotides. *Physical Review E*, 49(2):1685, 1994.

[73] Shaun Lovejoy and Daniel Schertzer. Haar wavelets, fluctuations and structure functions: convenient choices for geophysics. *Nonlinear Processes in Geophysics*, 19(5):513–527, 2012.

[74] Jianbo Gao, Jing Hu, and Wen-wen Tung. Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PloS One*, 6(9):e24331, 2011.

[75] B W Lindgren. *Statistical Theory*. MacMillan, New York, 1976.

[76] T Graves, RB Gramacy, CLE Franzke, and NW Watkins. Efficient bayesian inference for arfima processes. *Nonlinear Processes in Geophysics Discussions*, 2:573–618, 2015.

[77] Peter Whittle. Estimation and information in stationary time series. *Arkiv för matematik*, 2(5):423–434, 1953.

[78] Robert Fox and Murad S Taqqu. Large-sample properties of parameter estimates for strongly dependent stationary gaussian time series. *The Annals of Statistics*, pages 517–532, 1986.

[79] L Giraitis and D Surgailis. A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of whittle's estimate. *Probability Theory and Related Fields*, 86(1):87–104, 1990.

[80] Murad S Taqqu and Vadim Teverovsky. *A practical guide to heavy tails: Statistical techniques and applications*, chapter On estimating the intensity of long-range dependence in finite and infinite variance time series, pages 177–217. Birkhauser, Boston, 1998.

[81] Christian L E Franzke, Timothy Graves, Nicholas W Watkins, Robert B Gramacy, and Cecilia Hughes. Robustness of estimators of long-range dependence and self-similarity under non-gaussianity. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 370(1962):1250–1267, 2012.

[82] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[83] Jan Beran, R J Bhansali, and D Ocker. On unified model selection for stationary and nonstationary short-and long-memory autoregressive processes. *Biometrika*, 85(4):921–934, 1998.

[84] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[85] K P Burnham and D R Anderson. *Model selection and multimodel inference*. Springer-Verlag, New York, 2002.

[86] World Meteorological Organization. *Guide to Climatological Practices WMO-No. 100*. Number 100. 2011.

[87] Armin Bunde, Josef Ludescher, Christian LE Franzke, and Ulf Büntgen. How significant is west antarctic warming? *Nature Geoscience*, 7(4):246–247, 2014.

[88] H. W. Rust. *Detection of Long-Range Dependence – Applications in Climatology and Hydrology*. PhD thesis, University of Potsdam, Germany, 2007.

[89] William S Cleveland, Colin L Mallows, and Jean E McRae. Ats methods: Nonparametric regression for non-gaussian data. *Journal of the American Statistical Association*, 88(423):821–835, 1993.

[90] Engineering National Academies of Sciences and Medicine. *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*. The National Academies Press, Washington, DC, 2016.

[91] BD Santer, JJ Hnilo, TML Wigley, JS Boyle, C Doutriaux, M Fiorino, DE Parker, and KE Taylor. Uncertainties in observationally based estimates of temperature change in the free atmosphere. *Journal of Geophysical Research: Atmospheres*, 104(D6):6305–6333, 1999.

[92] Andrew J Majda and Boris Gershgorin. Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences*, 107(34):14958–14963, 2010.

[93] Robert B Cleveland, William S Cleveland, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3, 1990.

[94] Roland A Madden. A simple approximation for the variance of meterological time averages. *Journal of Applied Meteorology*, 18(5):703–706, 1979.

[95] H Jean Thiébaux and Francis W Zwiers. The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology*, 23(5):800–811, 1984.

[96] FW Zwiers and HJ Thiébaux. Statistical considerations for climate experiments. part i: Scalar tests. *Journal of Climate and Applied Meteorology*, 26(4):464–476, 1987.

[97] Regina Nuzzo. Statistical errors. *Nature*, 506(7487):150–152, 2014.

[98] William H Starbuck. 60th anniversary essay: How journals could improve research practices in social science. *Administrative Science Quarterly*, 61(2):165–183, 2016.

[99] Yoshikazu Hayashi. Confidence intervals of a climatic signal. *Journal of the Atmospheric Sciences*, 39(9):1895–1905, 1982.

[100] M Massah and H Kantz. Confidence intervals for time averages in the presence of long range correlations, a case study on earth surface temperature anomalies. *Geophysical Research Letters*, (10.1002/2016GL069555), 2016.

[101] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[102] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[103] Jan Beran. A test of location for data with slowly decaying serial correlations. *Biometrika*, 76(2):261–269, 1989.

[104] J Neyman and ES Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society*, 231:289–337, 1933.

[105] John Hinde. Choosing between non-nested models: a simulation approach. In *Advances in GLIM and Statistical Modelling*, pages 119–124. Springer, 1992.

[106] Jan Beran. A goodness-of-fit test for time series with long range dependence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 749–760, 1992.

[107] Aiko Voigt and Tiffany A Shaw. Circulation response to warming shaped by radiative changes of clouds and water vapour. *Nature Geoscience*, 8(2):102–106, 2015.

[108] United Nations Development Programme. *Fighting climate change: human solidarity in a divided world*. Springer, 2007.

[109] World Economic Forum Water Initiative. *Water security: the water-food-energy-climate nexus*. Island Press, 2012.

[110] FHS Chiew, NJ Potter, J Vaze, C Petheram, L Zhang, J Teng, and DA Post. Observed hydrologic non-stationarity in far south-eastern australia: implications for modelling and prediction. *Stochastic Environmental Research and Risk Assessment*, 28(1):3–15, 2014.

[111] Martin Hoerling, Jon Eischeid, Judith Perlwitz, Xiaowei Quan, Tao Zhang, and Philip Pegion. On the increased frequency of mediterranean drought. *Journal of Climate*, 25(6):2146–2161, 2012.

[112] Daniel R Cayan, Tapash Das, David W Pierce, Tim P Barnett, Mary Tyree, and Alexander Gershunov. Future dryness in the southwest us and the hydrology of the early 21st century drought. *Proceedings of the National Academy of Sciences*, 107(50):21271–21276, 2010.

[113] MS Morales, DA Christie, R Villalba, J Argollo, J Pacajes, JS Silva, CA Alvarez, JC Llancabure, and CC Soliz Gamboa. Precipitation changes in the south american altiplano since 1300 ad reconstructed by tree-rings. *Climate of the Past*, 8(2):653, 2012.

[114] Jian-Sheng Ye. Trend and variability of china's summer precipitation during 1955–2008. *International Journal of Climatology*, 34(3):559–566, 2014.

[115] Aiguo Dai. Drought under global warming: a review. *Wiley Interdisciplinary Reviews: Climate Change*, 2(1):45–65, 2011.

[116] Aiguo Dai. Increasing drought under global warming in observations and models. *Nature Climate Change*, 3(1):52–58, 2013.

[117] Jesse Kenyon and Gabriele C Hegerl. Influence of modes of climate variability on global precipitation extremes. *Journal of Climate*, 23(23):6248–6262, 2010.

[118] Ben D Santer, Michael F Wehner, TML Wigley, R Sausen, GA Meehl, KE Taylor, C Ammann, J Arblaster, WM Washington, JS Boyle, et al. Contributions of anthropogenic and natural forcing to recent tropopause height changes. *Science*, 301(5632):479–483, 2003.

[119] Dian J Seidel and William J Randel. Recent widening of the tropical belt: Evidence from tropopause observations. *Journal of Geophysical Research: Atmospheres*, 112(D20), 2007.

[120] Dian J Seidel, Qiang Fu, William J Randel, and Thomas J Reichler. Widening of the tropical belt in a changing climate. *Nature geoscience*, 1(1):21–24, 2008.

[121] Christopher Lucas, Bertrand Timbal, and Hanh Nguyen. The expanding tropics: a critical assessment of the observational and modeling studies. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1):89–112, 2014.

[122] Y Hu and Q Fu. Observed poleward expansion of the hadley circulation since 1979. *Atmospheric Chemistry and Physics*, 7(19):5229–5236, 2007.

[123] Jian Lu, Gabriel A Vecchi, and Thomas Reichler. Expansion of the hadley cell under global warming. *Geophysical Research Letters*, 34(6), 2007.

[124] Chia Chou, J David Neelin, Chao-An Chen, and Jien-Yi Tu. Evaluating the "rich-get-richer" mechanism in tropical precipitation change under global warming. *Journal of Climate*, 22(8):1982–2005, 2009.

[125] Jian Lu, Clara Deser, and Thomas Reichler. Cause of the widening of the tropical belt since 1958. *Geophysical Research Letters*, 36(3), 2009.

[126] Qiang Fu and Pu Lin. Poleward shift of subtropical jets inferred from satellite-observed lower-stratospheric temperatures. *Journal of climate*, 24(21):5597–5603, 2011.

[127] Thomas Reichler. Changes in the atmospheric circulation as indicator of climate change. In Trevor M Letcher, editor, *Climate Change: Observed Impacts on Planet Earth*, pages 145–164. Elsevier, Amsterdam, 2009.

[128] Robert J Allen, Steven C Sherwood, Joel R Norris, and Charles S Zender. Recent northern hemisphere tropical expansion primarily driven by black carbon and tropospheric ozone. *Nature*, 485(7398):350–354, 2012.

[129] Robert J Allen, Joel R Norris, and Mahesh Kovilakam. Influence of anthropogenic aerosols and the pacific decadal oscillation on tropical belt width. *Nature Geoscience*, 7(4):270–274, 2014.

[130] Damianos F Mantsis, Steven Sherwood, Robert Allen, and Lei Shi. Natural variations of tropical width and recent trends. *Geophysical Research Letters*, 44(8):3825–3832, 2017.

[131] Thomas Birner. Recent widening of the tropical belt from global tropopause statistics: Sensitivities. *Journal of Geophysical Research: Atmospheres*, 115(D23), 2010.

[132] Yongyun Hu, Chen Zhou, and Jiping Liu. Observational evidence for poleward expansion of the hadley circulation. *Advances in Atmospheric Sciences*, 28:33–44, 2011.

[133] Sean M Davis and Karen H Rosenlof. A multidiagnostic intercomparison of tropical-width time series using reanalyses and satellite observations. *Journal of Climate*, 25(4):1061–1078, 2012.

[134] Celeste M Johanson and Qiang Fu. Hadley cell widening: Model simulations versus observations. *Journal of Climate*, 22(10):2713–2725, 2009.

[135] Yongyun Hu, Lijun Tao, and Jiping Liu. Poleward expansion of the hadley circulation in cmip5 simulations. *Advances in Atmospheric Sciences*, 30:790–795, 2013.

[136] YP Zhou, Kuan-Man Xu, YC Sud, and AK Betts. Recent trends of the tropical hydrological cycle inferred from global precipitation climatology project and international satellite cloud climatology project data. *Journal of Geophysical Research: Atmospheres*, 116(D9), 2011.

[137] Richard P Allan and Brian J Soden. Atmospheric warming and the amplification of precipitation extremes. *Science*, 321(5895):1481–1484, 2008.

[138] K-M Lau and H-T Wu. Detecting trends in tropical rainfall characteristics, 1979–2003. *International Journal of Climatology*, 27(8):979–988, 2007.

[139] Luciano Telesca, Jorge O Pierini, and Beatrice Scian. Investigating the temporal variation of the scaling behavior in rainfall data measured in central argentina by means of detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 391(4):1553–1562, 2012.

[140] Ping Yang, Wei Hou, and Guo-Lin Feng. *Wuli Xuebao/Acta Physica Sinica*, volume 57, chapter Determining the threshold of extreme events with detrended fluctuation analysis, pages 5333–5342. 08 2008.

[141] Haibo Du, Zhengfang Wu, Shengwei Zong, Xiangjun Meng, and Lei Wang. Assessing the characteristics of extreme precipitation over northeast china using the multifractal detrended fluctuation analysis. *Journal of Geophysical Research: Atmospheres*, 118(12):6165–6174, 2013.

[142] Qi Zhang, Jiquan Zhang, Denghua Yan, and Yongfang Wang. Extreme precipitation events identified using detrended fluctuation analysis (dfa) in anhui, china. *Theoretical and applied climatology*, 117(1-2):169–174, 2014.

[143] Armin Bunde, Jan F Eichner, Jan W Kantelhardt, and Shlomo Havlin. Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Physical Review Letters*, 94(4):048701, 2005.

[144] George J Huffman, David T Bolvin, Eric J Nelkin, David B Wolff, Robert F Adler, Guojun Gu, Yang Hong, Kenneth P Bowman, and Erich F Stocker. The trmm multisatellite precipitation analysis (tmpa): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of hydrometeorology*, 8(1):38–55, 2007.

[145] Jianbo Gao, Yinhe Cao, Wen-wen Tung, and Jing Hu. *Multiscale analysis of complex time series: integration of chaos and random fractal theory, and beyond*. John Wiley & Sons, 2007.

[146] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[147] Saptarshi Guha, Ryan Hafen, Jeremiah Rounds, Jin Xia, Jianfu Li, Bowei Xi, and William S Cleveland. Large complex data: divide and recombine (d&r) with rhipe. *Stat*, 1(1):53–67, 2012.

[148] William S Cleveland and Ryan Hafen. Divide and recombine (d&r): Data science for large complex data. *Statistical Analysis and Data Mining*, 7(6):425–433, 2014.

[149] Tetsuzo Yasunari. The monsoon year—a new concept of the climatic year in the tropics. *Bulletin of the American Meteorological Society*, 72(9):1331–1338, 1991.

[150] David Molden. *Water for food, water for life: a comprehensive assessment of water management in agriculture*. Earthscan, 2007.

[151] Alex de Sherbinin. The biophysical and geographical correlates of child malnutrition in africa. *Population, Space and Place*, 17(1):27–46, 2011.

[152] Alex De Sherbinin, Leah K VanWey, Kendra McSweeney, Rimjhim Aggarwal, Alisson Barbieri, Sabine Henry, Lori M Hunter, Wayne Twine, and Robert Walker. Rural household demographics, livelihoods and the environment. *Global environmental change*, 18(1):38–53, 2008.