AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2014JD022506

Key Points:
- Spatial means and standard deviations of temperature anomalies are correlated
- Warming is more heterogeneous than cooling
- The Medieval Climate Anomaly is a period of increased spatial variability

Correspondence to:
M. P. Tingley,
mpt14@psu.edu

Citation:
Tingley, M. P., and P. Huybers (2015), Heterogeneous warming of Northern Hemisphere surface temperatures over the last 1200 years, *J. Geophys. Res. Atmos.*, *120*, 4040–4056, doi:10.1002/2014JD022506.

Received 29 AUG 2014
Accepted 10 APR 2015
Accepted article online 20 APR 2015
Published online 14 May 2015

Heterogeneous warming of Northern Hemisphere surface temperatures over the last 1200 years

Martin P. Tingley[1] **and Peter Huybers**[2]

[1]Department of Meteorology and Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, USA, [2]Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, USA

Abstract The relationship between the mean and spatial variability of Northern Hemisphere surface temperature anomalies over the last 1200 years is examined using instrumental and proxy records. Nonparametric statistical tests applied to 14 well-studied, annually resolved proxy records identify two centuries roughly spanning the Medieval Climate Anomaly as characterized by increased spatial variability relative to the preinstrumental baseline climate, whereas two centuries spanning the Little Ice Age are characterized by decreased spatial variability. Analysis of the instrumental record similarly indicates that the late and middle twentieth century warm intervals are generally associated with increased spatial variability. In both proxy and instrumental records an overall relationship between the first two moments is identified as a weak but significant positive correlation between time series of the spatial mean and spatial standard deviation of temperature anomalies, indicating that warm and cold anomalies are respectively associated with increased and reduced spatial variability. Insomuch as these historical patterns of relatively heterogeneous warming as compared with cooling hold, they suggest that future warming will feature increased regional variability.

1. Introduction

The temporal evolution of mean surface temperatures has been extensively studied using instrumental records and climate proxies [see, for example, *Masson-Delmotte et al.*, 2013, and references therein]. Far less attention has been given to the temporal evolution of spatial variability and its relationship to changes in the mean. Connecting local and global temperature changes, however, requires some understanding of how spatial variability relates to mean temperature. For example, a concurrent increase in both the mean and variance, as compared with an increase in the mean alone, results in a larger increase in the probability of crossing high-temperature thresholds [e.g., *Katz and Brown*, 1992; *Intergovernmental Panel on Climate Change*, 2007, Figure SPM.3]. The structure of spatial variations also determines the degree to which global changes can be inferred from point observations [e.g., *Tingley et al.*, 2012].

Studies of recent changes in distributions of instrumentally recorded surface temperatures have examined standard deviations through time, or after binning temperature anomalies in both space and time [e.g., *Hansen et al.*, 2012; *Huntingford et al.*, 2013], and then compared the resulting distributions across specific time intervals. Using this approach—which combines elements of both spatial and temporal variability—it has been difficult to identify changes in the distribution beyond those associated with shifts in the mean [*Rhines and Huybers*, 2013; *Huntingford et al.*, 2013; *Tingley and Huybers*, 2013].

An important point of comparison for the modern warming is the “Medieval Climate Anomaly” (MCA). The MCA has been widely described as an overall warming composed of a series of nonsynchronous regional temperature excursions [see *Diaz et al.*, 2011; *Masson-Delmotte et al.*, 2013, and references therein], implying an increase in spatial variability concurrent with increased mean temperatures. There is likewise ambiguity over the timing and spatial extent of the “Little Ice Age” (LIA) cooling [*Masson-Delmotte et al.*, 2013]. Insomuch as the direction, as well as the magnitude, of anomalies in the mean time series affects spatial variability, the MCA and LIA may feature different spatial standard deviation signatures. Also relevant are the conclusions of *Osborn and Briffa* [2006] that recent decades feature more spatially uniform warmth, as measured by threshold crossings, than any other interval in the last 1200 years. The degree to which the analysis of *Osborn and Briffa* [2006] allows for changes in mean to be disentangled from other changes in the distribution, however, is unclear.

Here we explore the temporal relationship between changes in spatial average temperature and the spatial variability about that average using both paleoclimate and instrumental records. Specifically, we construct statistical hypothesis tests to investigate if changes in both the spatial mean and standard deviation observed during periods we identify as the MCA and the LIA are significant with respect to the pretwentieth century climate record. We then summarize different intervals of both the proxy and instrumental data sets using two-moment “fingerprints” of the mean and standard deviation. Finally, we assess the significance of positive correlations between the mean and standard deviation observed in both the proxy and instrumental data sets.

2. Data and Methods

2.1. Instrumental Data

The instrumental temperature data are the 624 HadCRUT4 [*Morice et al.*, 2012] Northern Hemisphere time series with at least eight monthly values for each year from 1919 to 2013 (Figure 1). HadCRUT4 is a gridded data product derived from weather station and ship-based observations, and we use only those grid boxes with complete time series from 1919 to 2013 in order to avoid complications arising from changing data availability with time [*Tingley*, 2012]. To avoid the short 1961 – 1990 reference interval used to standardize the instrumental time series [*Brohan et al.*, 2006] from introducing spurious structures in the time series of standard deviations [*Tingley*, 2012], we remove the mean calculated over the entire 100 year interval from each time series. Each time series is then smoothed with a nine-point Hanning window.

The requirement of at least eight monthly values a year and the 95 year interval were chosen as a balance between length, completeness, and number of records. We restrict the analysis to the post – World War I interval as the years 1914 – 1918 feature uncertainty intervals for the standard deviation estimates that are up to 3 times as wide as those for neighboring years outside of World War I. The increased uncertainty is likely due to the decrease in the number of observations entering the HadCRUT4 averaging scheme [see *Morice et al.*, 2012, Figure 5]. Although World War II also featured reductions in the observational network, we detect no anomalous behavior in time series of means or standard deviations during this interval, perhaps because the baseline network was sufficiently dense during that interval.

2.2. Proxy Data

The proxy data consist of 12 of the records used by *Osborn and Briffa* [2006], as well as the Shihua cave record and Indigirka tree ring record used by *Moberg et al.* [2005]. These last two records are the only two annually resolved records used by *Moberg et al.* [2005] that are not collocated with records used by *Osborn and Briffa* [2006] and are added to increase sample size and spatial coverage (Figure 1). Detailed descriptions of the data series are available in the supplements to *Osborn and Briffa* [2006] and *Moberg et al.* [2005]. We exclude the Chesapeake Bay Mg/Ca [*Cronin et al.*, 2003] and East Asian regional multiproxy [*Yang et al.*, 2002] records used in *Osborn and Briffa* [2006], as the former features a low sampling rate prior to about 1600, and the latter appears to be heavily smoothed, having less than 5% of its spectral power at periods shorter than 50 years (Figure 1). We retain the smoother Mongolian composite [*D’Arrigo et al.*, 2001] used by *Osborn and Briffa* [2006] as spectral analysis indicates that it preserves decadal variability (approximately 50% of spectral power lies in periods between 10 and 100 years), and because we standardize the low-frequency components of all series (see below). For the same reasons, we include the documentary temperature record from the Netherlands and Belgium [*van Engelen et al.*, 2001], despite it featuring more high-frequency variability than the other records.

Prior to standardization, the proxy records are each smoothed by a 19 point Hanning window, in order to focus on the lower frequency fluctuations that are well resolved across all proxy records in this collection. This longer smoothing window is useful for analysis of the proxy records because they are noisier than the instrumental observations. Inasmuch as low-frequency temperature anomalies have a larger spatial extent, they will more readily be reconstructed from a sparse proxy set. Following *Osborn and Briffa* [2006], the smoothed proxy records are standardized by first setting the mean and variance of all complete records to zero and one, respectively. Then, in order of decreasing length of the records, the mean and variance of each record is set to match the mean and variance of the already standardized records over the period of overlap (Figure 1).

The proxies feature a positive association with annual temperatures. During the 1919 – 1990 overlap between the smoothed proxy and instrumental data sets, the correlation between the proxy and instrumental mean time series is 0.79, indicating that the proxies are capturing the rapid temperature rise over the twentieth century. Some fraction of the variability across the proxy records at any fixed time is a result of observational error and the different relationships between these proxy types and the actual temperature anomalies.

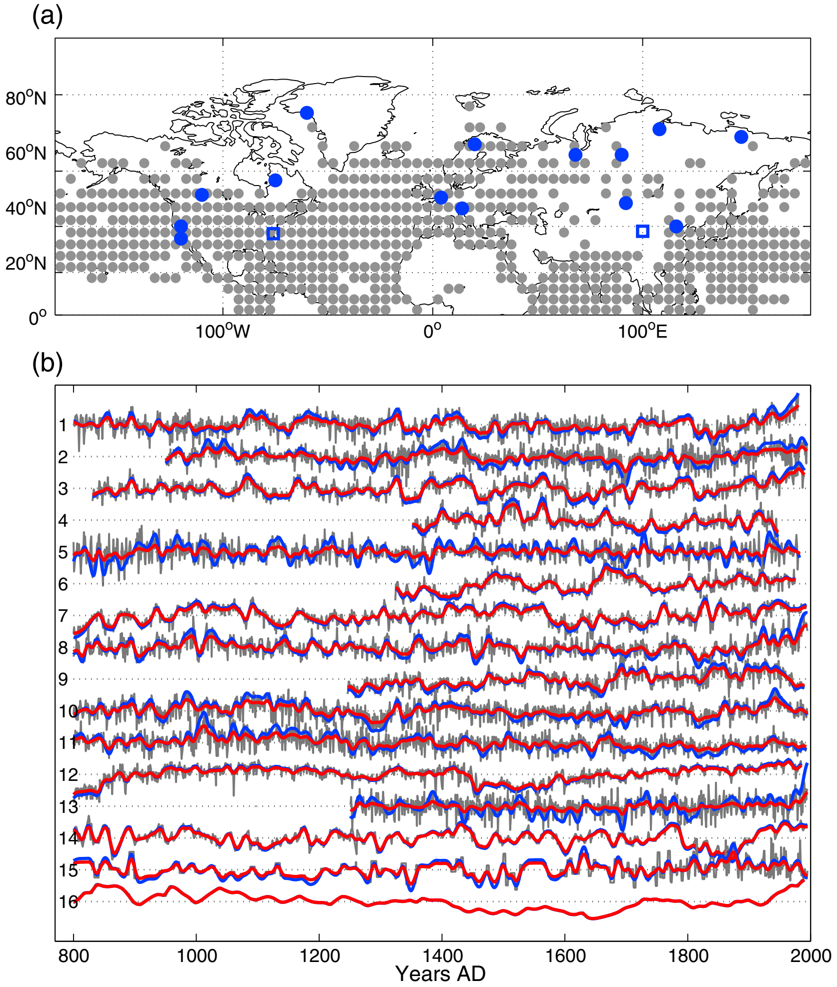

Figure 1. (a) Gray circles: locations of the 624 HadCRUT4 instrumental time series. Blue circles: locations of the proxy records. Blue squares: locations of the two records used by *Osborn and Briffa* [2006] but excluded here. (b) Proxy time series. Light grey: standardized raw data. Red: standardized then smoothed with a 19 point Hanning window. Blue: smoothed then standardized. The bottom two panels in Figure 1b are the two records used by [*Osborn and Briffa*, 2006] but excluded from this study. See *Osborn and Briffa* [2006] and *Moberg et al.* [2005] for references to the data sets. The records are (1) western USA (principal component of tree ring chronologies); (2) Southwest Canada (Icefields; RCS (Regional Curve Standardization) tree ring chronology); (3) western USA (Boreal/Upperwright; two RCS tree ring chronologies); (4) Northeastern Canada (Quebec; RCS chronology); (5) West Greenland (regional; composite of δ^{18}O ice core records); (6) Austria (Tirol; RCS tree ring chronology); (7) Northern Sweden (Tornetrask; RCS tree ring chronology); (8) Northwest Russia (Yamal; RCS tree ring chronology); (9) Northwest Russia (Mangazeja; RCS tree ring chronology); (10) Northern Russia (Taimyr; RCS tree ring chronology); (11) Western Russia (Indigirka; tree ring chronology); (12) China (Beijing temperature reconstruction via stalagmite layer thickness); (13) Netherlands and Belgium (regional; documentary); (14) Mongolia (regional; composite of tree ring chronologies); (15) eastern USA (Chesapeake Bay; Mg/Ca in fossil shells); (16) East Asia (regional; composite of multiple proxy types).

We assume that this noise component of the proxy records (see equations (A1) and (A2)) is unstructured in time, noting the absence of an overall trend in the time series of spatial standard deviations and that there are no sudden changes in the mean or standard deviation time series, or their uncertainties, as the number of records changes (Figure 2). Observed structure is thus interpreted as changes in the spatial standard deviation of temperature anomalies, which we expect to be time variable.

We confine the proxy analysis to 831–1990, during which there are at least nine proxy observations a year, judging that with fewer records results become more difficult to interpret. Our qualitative conclusions remain unchanged, however, if the analysis is confined to the six proxy records that are complete over the 831–1990 interval. In what follows, we also perform analyses on the 1896–1990 interval of the proxy data for comparison

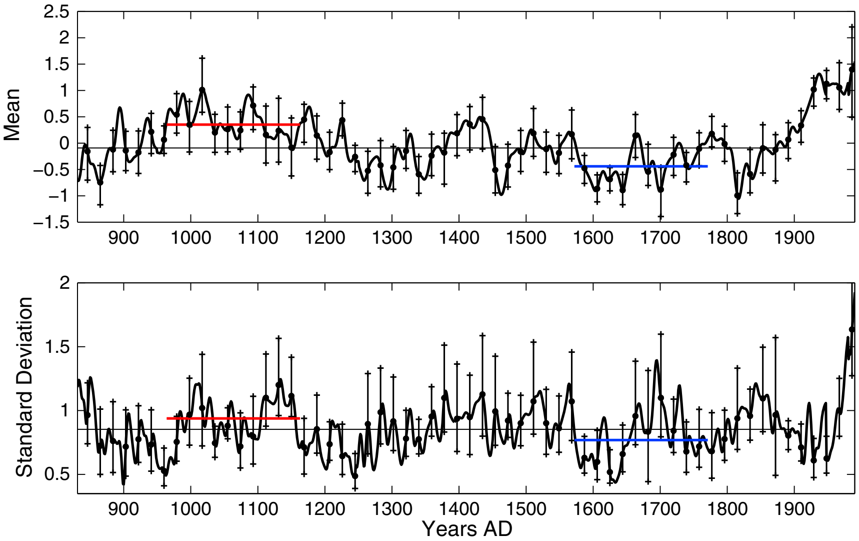

Figure 2. (top) Mean and (bottom) standard deviation across the smoothed proxy data. Every 19 years 90% confidence intervals are shown. Gray lines: median values of the mean and standard deviations in the 831–1895 interval. Red lines: median values within the 964–1163 MCA. Blue lines: median values within the 1572–1771 LIA.

with results from the instrumental data. Although the comparison between the 1896–1990 interval of the proxy data and the 1919–2013 interval covered by the instrumental data is imperfect, both of these 95 year intervals are dominated by rapid warming, and confining the proxies to the 72 year overlapping period of 1919–1990 results in uncertainties, particularly when estimating correlations, that are so large as to preclude meaningful comparisons.

2.3. Uncertainty Estimates for the Mean and Standard Deviation

We calculate 90% confidence intervals for spatial means and standard deviations of the smoothed proxy (Figure 2) and instrumental (Figure 3) data sets using the bias corrected-accelerated (BCa) bootstrap [*Efron and Tibshirani*, 1986]. We counteract the effect of spatial correlation by resampling (with replacement) 50 of the 624 available instrumental anomalies at each iteration of the bootstrap (see Appendix A). The simplicity of the analysis comes at the cost of collapsing spatial patterns of variability onto scalar quantities but allows for simple tests of past changes in the temperature distribution.

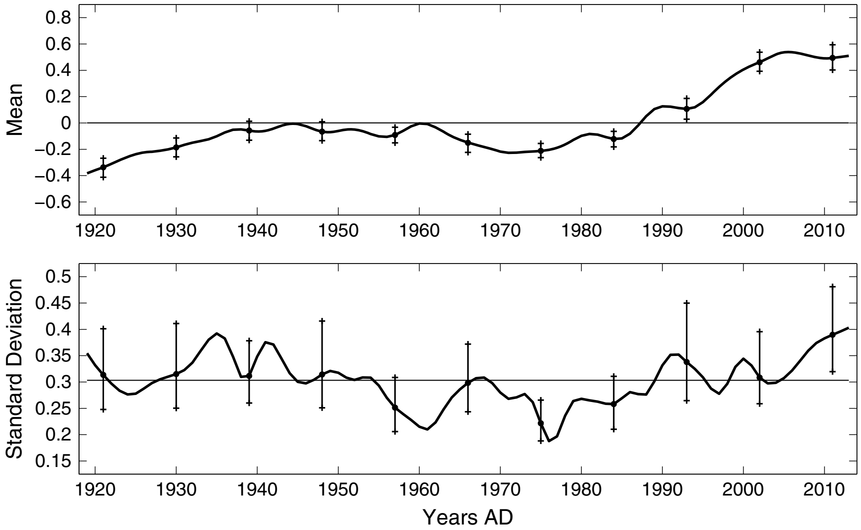

Figure 3. (top) Mean and (bottom) standard deviation for the smoothed instrumental data. Every 9 years 90% confidence intervals are shown.

2.4. Statistical Tests and Null Hypotheses

We use the Pearson product-moment correlation [*Zar*, 1999] to investigate the general relationship between the mean and standard deviation, and the rank-sum Mann-Whitney *U* statistic [*Zar*, 1999] to test for differences between time intervals:

$$U = n_2 n_1 + \frac{n_2(n_2+1)}{2} - R_2, \tag{1}$$

where n_1 is the number of years in the first interval, n_2 the number of years in the second, and R_2 is the sum of the ranks of the means or standard deviations in the second. *U* is invariant under monotonic transformations of the data and is therefore identical for time series of variances or standard deviations. The Mann-Whitney test is a nonparametric procedure for testing differences in median values between populations, and as the standard assumption of independent samples is clearly violated in our case (Figures 2 and 3), we use Monte Carlo procedures to establish significance.

Following *Osborn and Briffa* [2006], we test the significance of sample statistics against a *rotational null* distribution, formed by repeatedly randomizing the starting points of the records in each data set and wrapping the portion of the rotated record that extends beyond the endpoint to fill in its beginning. For the proxy data, the rotation procedure is confined to the section of each record for which there are observations. The rotational null is used to test features of the data against surrogate data sets with similar time series properties but with no spatial correlation. We also test significance against a null formed by simulating from AR(1) processes fit to each time series. For tests of the difference in median values between two time intervals, the AR(1) null is easier to reject than the rotational null, and both are more difficult to reject than the standard Mann-Whitney test. Likewise, for tests of the significance of correlations, the rotational null is more difficult to reject than the AR(1) null. Throughout, we report one-sided *p* values, as the literature suggests a priori that the medieval period is both warm and variable and therefore that any correlation between the mean and standard deviation should be positive.

As a complement to testing the observed correlation coefficients against the null distributions, we also estimate the uncertainty in sample correlation values by bootstrapping the pairs of mean and standard deviation estimates [*Efron and Tibshirani*, 1986], accounting for the serial correlation of the data sets (see Appendix A). The confidence intervals and hypothesis tests do not give equivalent information, as we find below that the rotational null is centered on a positive value, meaning that the confidence interval may not contain zero despite a failure to reject the rotational null.

3. Results

3.1. Mean and Standard Deviation Time Series

The time series of instrumental means (Figure 3) is in agreement with results from the Intergovernmental Panel on Climate Change Fifth Assessment Report [*Hartmann et al.*, 2013]: temperatures rise to a broad local maximum near the middle of the twentieth century, dip to a local minimum at about 1970, and then steadily rise, with the warmest temperatures clustered over the last decade. The temporal variability of the point estimates is large relative to the uncertainty at each year. The standard deviations are harder to interpret (Figure 3) as the uncertainty in each estimate is about as large as the temporal range of the estimates. Qualitatively, the standard deviation time series exhibits some features that are similar to the mean time series: the standard deviations are generally large between 1930 and 1945, fall to a local minimum between 1960 and 1980, and then increase to a level on par with the 1930s.

For a given change in the mean temperature series, the increase in the spatial standard deviation will be largest if the change in the mean is caused by a change in a single constituent time series. The standard deviation time series therefore indicates the spatial heterogeneity of temperature changes, as the smaller the number of records responsible for a change in the mean, the larger the increase in the standard deviation.

To investigate if years that feature high values of the standard deviation feature a common spatial signature, and likewise for years that feature small standard deviations, we calculate the temporal average of the spatial deviations across the 24 years with highest, or lowest, standard deviations, removing the spatial mean from each year (Figure 4). The two resulting patterns of average spatial deviations feature continental-scale variations, and the correlation between the two is −0.74. The pattern of average spatial deviations for the high standard deviation years features generally positive values over land and negative values over most ocean basins, with the North Atlantic being a notable exception; the pattern for the low standard deviation years is

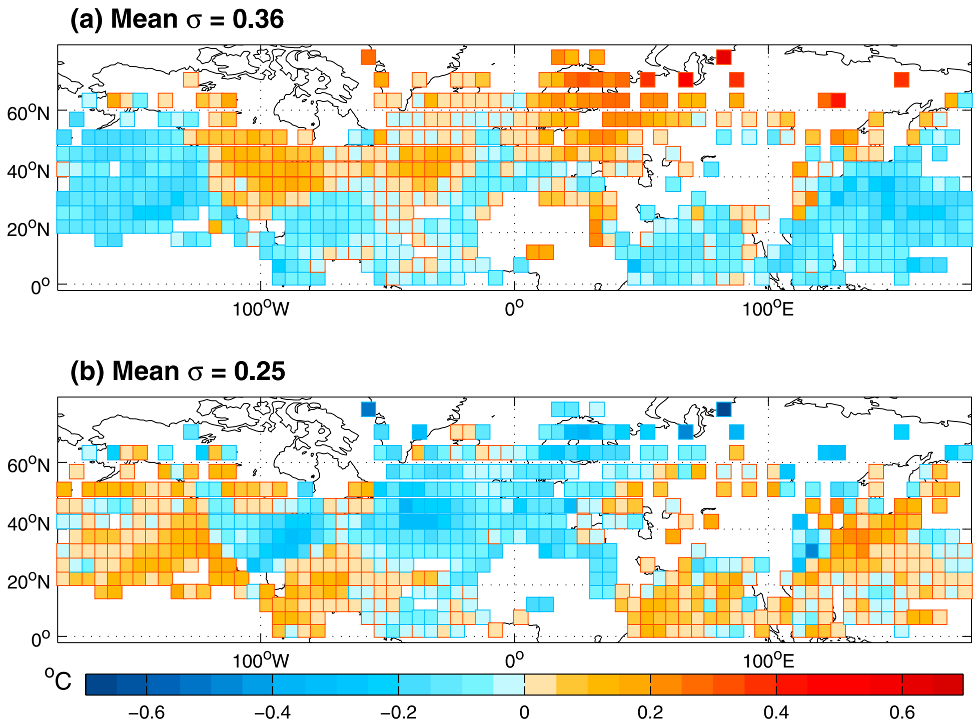

Figure 4. Spatial distributions of HadCRUT4 [*Morice et al.*, 2012] temperature deviations averaged over (a) the 24 years featuring the largest standard deviations and (b) the 24 years featuring the smallest standard deviations. Panels are on a common temperature scale, and the spatial mean has been removed in each case.

largely the reverse. These features suggest that the positive relationship between mean and standard deviation identified below is controlled by consistent, and largely opposite, regional-scale temperature patterns that feature high or low variability.

A second notable feature is that the largest magnitude mean spatial deviations for the high standard deviation years are positive and clustered at high latitudes, whereas those for the low standard deviation years are negative but clustered in the same region (Figure 4). These features suggest an asymmetric Arctic amplification [cf. *Collins et al.*, 2013], whereby domain-wide warming results in larger magnitude warm anomalies in the Arctic than the corresponding cool anomalies under domain-wide cooling.

As the proxies are terrestrial, they are unable to detect the land-sea contrasts evident in the patterns of average spatial deviations from the instrumental data (Figure 4). However, there is good proxy coverage in the high-latitude region that features the largest magnitude average spatial deviations of both signs, according to the instrumental data (Figure 4) and which is therefore most important in determining the spatial standard deviation of temperature anomalies.

The mean time series for the smoothed proxy anomalies features a broad maximum in the medieval period, an uneven cooling to a minimum at about 1600, and a steady rise after about 1800 (Figure 2), broadly consistent with other reconstructions [e.g., *Masson-Delmotte et al.*, 2013]. Confidence intervals are narrow relative to the changes in the mean, and the confidence intervals for warm periods do not overlap with those for cool periods. The standard deviation time series of the proxy records features greater variability than the mean time series, and the confidence intervals are wider relative to changes in the point estimates (Figure 2). As with the instrumental data, there are qualitative similarities between the means and standard deviations, with the latter rising over the most recent century and featuring high values concurrent with a warm interval centered on about 1000 A.D.

In what follows, we test and quantify the relationship between the mean and standard deviation time series, for both the instrumental and proxy data sets.

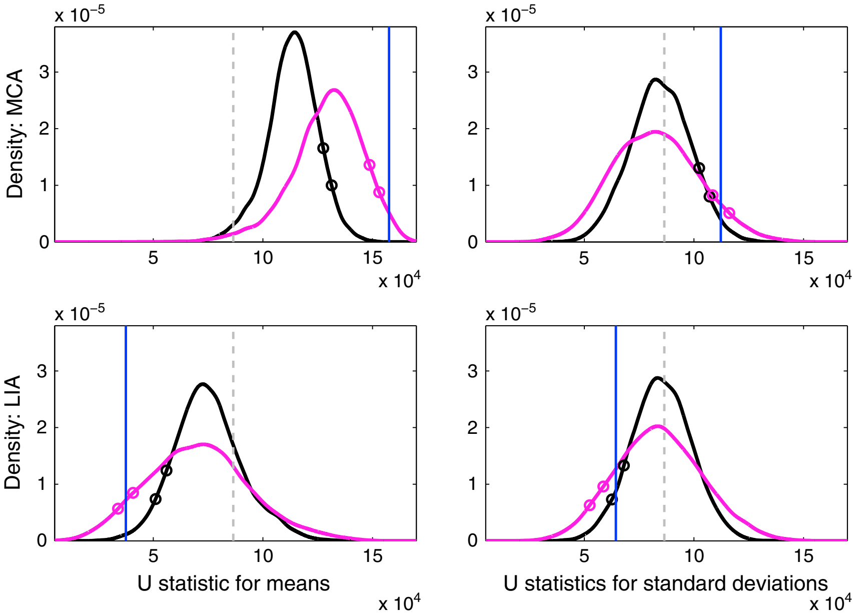

Figure 5. *U* statistics and null distributions for comparing the (left column) spatial mean and (right column) standard deviation of the smoothed proxy data during the (top row) 964–1163 MCA and (bottom row) 1572–1771 LIA to pre-1896 values outside of each interval. Vertical grey lines: expected values of the *U* statistics under the assumptions of the Mann-Whitney test. Vertical blue lines: data values. Black: kernel density estimates of the AR(1) null distributions. Purple: kernel density estimates of the rotational null distributions. One-sided 90% and 95% critical values according to each null are marked with circles on the density estimates. See Table 1 for numerical results.

3.2. Means and Standard Deviations During the MCA and LIA

For the proxies analyzed here, 964–1163 A.D. maximizes the Mann-Whitney *U* for comparing the mean time series within a 200 year interval to values outside that interval but prior to 1896. The interval defined as the MCA differs across studies, with, for example, *Diaz et al.* [2011] using a 950–1400 A.D. interval and *Masson-Delmotte et al.* [2013] using 950–1250 A.D. We define the MCA as the shorter 964–1163 A.D. interval to focus results on the peak warming as recorded by the proxies used here.

The 964–1163 A.D. MCA features a median value that is 0.52 standardized proxy units larger than the median of pre-1896 values that lie outside of the MCA. A test of the rarity of a positive anomaly of this magnitude must account for the fact that the interval is identified on the basis of being warm. For each Monte Carlo iteration, according to each null hypothesis, we therefore select the 200 year interval that maximizes the *U* statistic for the mean time series (Figure 5). The *p* values for testing against the rotational and AR(1) null hypotheses are 0.02 and $<$ 0.001, respectively, indicating that the 964–1163 period is anomalously warm with respect to the preinstrumental baseline, even accounting for the fact that we are a priori selecting the warmest interval (Table 1).

Using the same procedure of selecting the warmest interval from each null realization, but now computing the standard deviation, indicates that the observed MCA increase in standard deviation of 0.11 units is likewise significant (Figure 5). The *p* values against the rotational and AR(1) nulls are 0.07 and 0.02, respectively (Table 1). There is evidence, with the level of significance lower than for tests of the mean and dependent on the null, that the MCA was a period of anomalously large spatial variability. These results provide quantitative evidence for previous inferences that the MCA is both warm and variable [*Hughes and Diaz*, 1994; *Crowley and Lowery*, 2000; *Esper et al.*, 2002; *Bradley et al.*, 2003; *D'Arrigo et al.*, 2006; *Guiot et al.*, 2010].

We identify the 1572–1771 interval as the LIA, as it is the 200 year pre-1896 interval that minimizes the *U* statistic. In contrast to the MCA, the LIA is characterized by a decrease in the spatial mean and standard deviation (Figures 2 and 5). The *p* values for the observed decrease in mean of 0.41 during the LIA are 0.07 and 0.006 against the rotational and AR(1) nulls, respectively, and the corresponding *p* values for the observed decrease in standard deviation of 0.10 units are 0.16 and 0.06 (Table 1). Although the *p* values are larger than those for the MCA, there is a weak suggestion of decreased variability over the LIA, and more significant

Table 1. *U* Statistics and Associated One-Sided *p* Values for Comparing Spatial Means and Standard Deviations of the Smoothed Proxy Data, During the 964–1163 MCA and the 1572–1771 LIA Intervals, to Pre-1896 Values Outside of These Respective Intervals[a]

			One-Sided *p* Values	
	Anomally	*U*	AR(1)	Rotational
MCA mean	0.52	157573	$<$ **0.001**	**0.02**
MCA standard deviation	0.11	112271	**0.02**	*0.07*
LIA mean	−0.41	37511	**0.006**	*0.07*
LIA standard deviation	−0.10	64384	*0.06*	0.16

[a] Anomaly refers to the difference in median values of the mean or standard deviation within the MCA or LIA as compared with preinstrumental values outside of each interval. The *p* values smaller than 0.05 are in bold, while *p* values between 0.1 and 0.05 are in italics. Under the standard assumptions of the Mann-Whitney test, the expected value of *U* for this sample size is 87,600, and the 5% and 95% critical values are 81,078 and 94,122.

evidence (with level dependent on the null) that the LIA is anomalously cool, even when accounting for the fact that it is selected as a cool interval.

Reconstructions of millennial-scale temperatures are generally displayed as temperature anomalies with respect to a mean value calculated over a subset of the instrumental interval (1960–1990 in *Jansen et al.* [2007]; 1881–1980 in *Masson-Delmotte et al.* [2013]). "Spaghetti plots" of different reconstructions tend to feature relatively good agreement during the medieval period, and more divergent behavior during the cooler 1600s [*Jansen et al.*, 2007, Figure 6.10; *Masson-Delmotte et al.*, 2013, Figure 5.7]. This feature, seemingly at odds with our inference of increased variability during the MCA and decreased variability during the LIA, is a result of the linear transformation of proxy units into temperature anomaly units. As the medieval period featured temperatures on par (depending on the location) with those in the midtwentieth century, the additive constant in the regression that results from inverting a linear forward model for the proxy, $T = \frac{1}{\beta} \cdot (P - \alpha)$, [cf. *Li et al.*, 2010; *Cressie and Tingley*, 2010; *Tingley and Huybers*, 2010a, 2010b; *Christiansen*, 2011; *Tingley and Li*, 2012], sets the anomalies in proxy units to nearly zero. Estimates of temperature anomalies are then insensitive to multiplication by what are often uncertain slope parameters. In short, compilations of reconstructions will naturally feature greater agreement over a short, common reference interval, and during epochs with values similar to those during the reference interval; see *Tingley* [2012] for further discussion.

3.3. Two-Moment Fingerprints of Distinct Time Intervals

We summarize distinct intervals of both the proxy and instrumental records by plotting two-moment fingerprints composed of average values of the mean and standard deviation time series (Figure 6). Uncertainties are calculated by scaling the average widths of the corresponding 90% confidence intervals over various time periods by the square root of the temporal degrees of freedom, estimated as the length of the time period divided by half the length of the smoothing window (Appendix A).

For the proxy data, we consider the 964–1163 MCA, the 1572–1771 LIA, and the 1896–1990 modern intervals (Figure 6). The LIA can be differentiated from the MCA on the basis of both moments, as the former is cooler and less variable than the latter. In contrast, the 1896–1990 interval is much warmer than the MCA, but there is no evidence that it is more or less variable (Figure 6). As the dramatic increase in temperatures over the last 25 years is not covered by these proxies, the comparison is primarily between the MCA and the midtwentieth century warming.

Osborn and Briffa [2006] point to the significance of the geographic extent of the twentieth century warming, suggesting it can be differentiated from other warming intervals by it spatial homogeneity. In contrast, we conclude that it is primarily changes in the mean, not the standard deviation, that drive this distinction (Figure 6).

For the instrumental data, we consider two-moment fingerprints for the periods 1937–1962, 1963–1986, and 1987–2013 (Figure 6). These intervals, defined by intersections of the mean time series with its value in 1963, correspond to the midtwentieth century warming, the cooling during the 1970s, and the more recent warming (Figure 3); results are not sensitive to other choices of interval boundaries that still generally

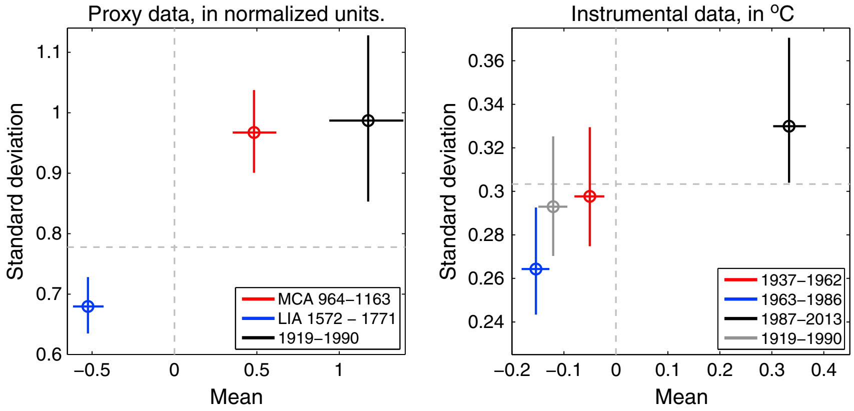

Figure 6. Averages of means and standard deviations, with associated uncertainties, over different intervals of the (left) smoothed proxy and (right) smoothed instrumental data sets. Dashed lines show averages over the entire spans of the data sets.

encompass these three anomalies. The three nonoverlapping periods are distinct from one another on the basis of the mean, and the standard deviations feature wide, overlapping uncertainties. The overall pattern, however, is similar to that for the proxies, with point estimates of the standard deviation monotonically increasing with increasing means. Instrumental results are likewise qualitatively unchanged when limiting the data to land areas, and when considering growing season, rather than annual, averages to better match the tree ring-dominated proxy data set.

Although the proxies are not in temperature units, results can be compared to the instrumental analysis by exploiting the fact that the two data sets overlap for the years 1919–1990 (Figure 6). According to the proxy analysis, the MCA was cooler than the 1919–1990 period, but the two intervals cannot be distinguished based on their respective spatial standard deviations. From the instrumental data, 1935–1962 period was similar in temperature or warmer than the longer 1919–1990 period, while 1987–2013 was substantially warmer. In the two dimensional space spanned by the mean and standard deviation, the MCA is closer to the midtwentieth century warming than the late twentieth century warming. The distinctions between time intervals are primarily driven by differences in mean values, but we stress that the pattern of larger standard deviation estimates accompanying larger mean values is consistent across both the proxy and instrumental data sets.

Several caveats with respect to the proxy analysis are worth noting. First, the MCA is a broad maximum in mean surface temperatures, whereas the modern interval features a more continuous rise in temperatures. Second, observations suggest that the statistical relation between tree rings and temperatures has weakened over the last few decades [*Jacoby and D'Arrigo*, 1995; *Briffa et al.*, 1998; *D'Arrigo et al.*, 2006]. It is thus possible that tree rings similarly became less sensitive to temperatures during previous warm intervals, biasing reconstructions. Provided such biases occur uniformly with warm temperatures, however, our conclusions are robust to this effect as the proxies would be biased low during both the MCA and the twentieth century, and we identify both intervals as warm. Also worth noting is evidence that the divergence between tree ring reconstructions and temperature is related to global dimming [*Stine and Huybers*, 2014; *Tingley et al.*, 2014] and is likely an exclusively twentieth century phenomena. This evidence suggests that the tree ring records are, in fact, not biased during the MCA. Given that we identify the modern interval as warmer than the MCA but cannot distinguish between these two warm intervals on the basis of their standard deviations (Figure 6), our results are robust to divergence resulting in the proxies underestimating mean temperature during the modern interval.

Finally, *Esper and Frank* [2009] suggest that reports of increased spatial variability during the MCA may be the result of data uncertainty and low replication for tree ring records. Here we find evidence for increased variability in both the MCA and the modern warm intervals and find the same pattern of increased warming accompanying larger mean values in the instrumental record. In addition, results are consistent, though more uncertain, when limiting the analysis to those six proxy records that are complete over 831–1990: both the modern interval and the MCA are warmer and more variable than the LIA, the modern interval is warmer than the MCA, but the two warm intervals cannot be distinguished on the basis of their standard deviations.

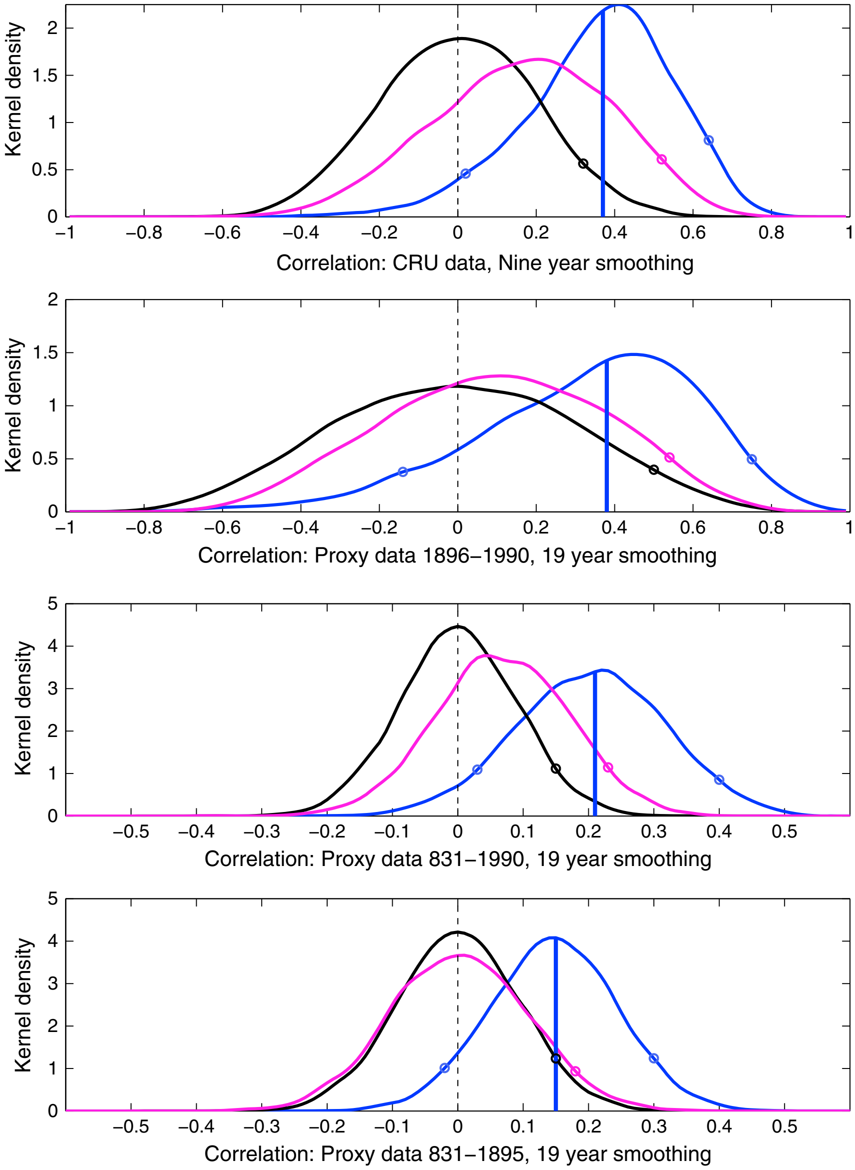

Figure 7. Correlations between the mean and standard deviation of the smoothed instrumental data (first panel) and three different intervals interval of the smoothed proxy data (1896–1990, 831–1990, and 831–1895). Blue vertical lines: observed correlations. Blue curves: kernel density estimate of the distribution of bootstrapped correlations. Limits of 90% confidence intervals are indicated with blue circles on the density estimates. Black and purple curves: kernel density estimates of the AR(1) and rotational null distributions, respectively, with upper 95% critical values marked with circles. Thin dashed lines at zero serve as references. Note that the *x* axes extend to ±1 in the first and second panels and to ±0.6 in the third and fourth panels. Numerical results are presented in Table 2.

3.4. Correlations Between Means and Standard Deviations

Graphical results for both data sets suggest that increases in mean and standard deviation occur in tandem (Figures 2, 3, and 6). Formal statistical tests confirm that the positive correlations between these two moments is significant, at least according to the AR(1) null. For the smoothed instrumental data, the Pearson correlation of 0.39 between the mean and standard deviation is significant at the 0.05 level according to the AR(1) null, and the 90% bootstrap confidence interval does not contain zero (Figure 7 and Table 2).

Over the 831–1990 interval spanned by the proxy data, the Pearson correlation is 0.21; *p* values against the AR(1) and rotational nulls are 0.01 and 0.08, respectively; and the 90% confidence interval does not cover zero (Figure 7 and Table 2). Despite the lower correlation, the longer time span covered by the proxy records leads to results that are significant at the 0.10 level according to the stricter of the two nulls.

Table 2. Correlations Between Estimates of the Mean and Standard Deviation From the Smoothed Proxy and Instrumental Data[a]

	Correlation	90% Confidence Interval	AR(1) p Value	Rotational p Value	N_{eq}
Instrumental (1919–2013)	0.39	**[0.05, 0.65]**	**0.03**	0.22	21
Proxy (1896–1990)	0.38	[−0.14, 0.76]	0.11	0.17	10
Proxy (831–1895)	0.15	[−0.02, 0.30]	*0.06*	*0.09*	112
Proxy (831–1990)	0.21	**[0.04, 0.40]**	**0.01**	*0.08*	122

[a] Also shown are 90% BCa bootstrap confidence intervals, resampling N_{eq} of the N pairs of means and standard deviations; one-sided p values against the AR(1) and rotational null hypotheses; and estimates of N_{eq}, the number of independent pairs of mean and standard deviation values. The p values smaller than 0.05 and 90% confidence intervals that do not cover zero are in bold, while p values between 0.1 and 0.05 are in italics.

We repeat the tests on the 831–1895 and 1896–1990 intervals of the proxy data, confining the construction of the rotational null to the target interval in each case. The sample correlation of 0.15 over the pre-1896 interval is on the margin of significance—p values against the AR(1) and rotational nulls are 0.06 and 0.09, respectively, and the 90% confidence interval extends down to −0.02. Although the correlation over the 1896–1990 interval is not significant at the 0.10 level (Figure 7 and Table 2), the higher value of 0.39 suggests that the proxy records are replicating the behavior of the instrumental records, despite the reduction in degrees of freedom precluding a significant result. In addition, inclusion of the 1896–1990 interval increases the correlation over the 831–1990 interval, as compared with the 831–1895 interval, and increases the significance against both nulls.

The correlations generally increase with greater smoothing, suggesting that the association between the mean level and the variability is primarily a low-frequency phenomenon. As an example, applying a 39 (as opposed to 19) year smoothing window to the proxies leads to a correlation of 0.24 over the 831–1990 period, while applying a 19 (as opposed to 9) year smoothing window to the instrumental series increases the correlation to 0.66. As the longer smoothing windows reduce the equivalent number of independent data points, however, establishing significance is more difficult. Results are also similar when using Spearman's ranked correlation, as opposed to the Pearson product-moment correlation, with significance at the 0.05 level against the AR(1) null for the instrumental data, and for the proxy data over the 831–1990 interval.

The rotational null distributions are centered on positive correlations if data from the most recent century are included in the analysis (Figure 7), indicating that the zero-centered AR(1) null with normal errors cannot account for qualitatively important aspects of surface temperature data sets over this period. As the rotational null removes all effects of spatial correlation, a positive association between the mean and standard deviation is found to be an inherent feature of the temperature anomaly distributions themselves. A likely explanation is that nonnormality in the temperature distribution, most prominent over the twentieth century, results in greater probability of very warm, rather than very cold, temperature excursions. Indeed, a positively skewed distribution would produce positive correlations between the first two moments even if temperature anomalies at all locations and times were independent of one another [cf. *Huybers et al.*, 2014]. Consistent with this possible explanation, the sample skewness of the instrumental data (1919–2013) is +0.37 and that for the proxy data (831–1990) is +0.12.

4. Discussion and Conclusion

A two-moment analysis of spatial means and standard deviations indicates that the MCA is both warm and variable with respect to a pretwentieth century baseline. The relative warmth of the Medieval period has been discussed [e.g., *Lamb*, 1965; *Hughes and Diaz*, 1994; *Jansen et al.*, 2007; *Mann et al.*, 2009; *Ljungqvist*, 2010; *Masson-Delmotte et al.*, 2013] and tested on individual records [*Cronin et al.*, 2003], whereas the increased spatial variability of the Medieval period has been discussed in numerous studies [*Hughes and Diaz*, 1994; *Crowley and Lowery*, 2000; *Esper et al.*, 2002; *Bradley et al.*, 2003; *D'Arrigo et al.*, 2006; *Guiot et al.*, 2010; *Diaz et al.*, 2011; *Masson-Delmotte et al.*, 2013] but not previously tested. It follows that because the Medieval period features increased variability, the definition of a Medieval Climate Anomaly will be sensitive to the particular suite of proxy records under consideration [*Crowley and Lowery*, 2000; *D'Arrigo et al.*, 2006; *Guiot et al.*, 2010; *Diaz et al.*, 2011; *Masson-Delmotte et al.*, 2013; *Lamb*, 1965; *Hughes and Diaz*, 1994; *Huang et al.*, 1997; *Crowley* *and Lowery*, 2000; *Broeker*, 2001; *Esper et al.*, 2002; *Goose et al.*, 2005; *Osborn and Briffa*, 2006; *Goose et al.*, 2006; *D'Arrigo et al.*, 2006].

Previous studies, including our own, were unable to reject a null hypothesis of temporally constant variability for surface temperature data [*Rahmstorf and Coumou*, 2011; *Hansen et al.*, 2012; *Rhines and Huybers*, 2013; *Huntingford et al.*, 2013; *Tingley and Huybers*, 2013]. We note that these studies consider temporal or spatiotemporal variability, and generally compare standard deviations between distinct time intervals. Here in contrast, we consider purely spatial variability calculated from temporally smoothed series and find evidence of differences in standard deviations between distinct time intervals. A two-moment fingerprint analysis of different time intervals (Figure 6) reveals that the MCA and modern period were both warmer and more variable than the LIA. Likewise, two-moment fingerprints of different interval periods covered by the instrumental data suggest that warm intervals generally feature greater variability, though differences in variability are not individually significant.

Positive associations between changes in the mean and variability are confirmed by an analysis of the temporal relationship between the spatial mean and spatial standard deviation. There is a positive, statistically significant correlation of 0.39 between the mean and standard deviation of the instrumental data set. For the proxies, the correlation is positive and strongest over the most recent century, while significance is dependent on the null hypothesis and analysis time frame (Table 2). On balance, we therefore expect warming to feature a more heterogeneous distribution of anomalies than cooling. Results of any paleoclimatic analysis are dependent on the particular records used in the analysis, and our findings suggest that estimates of spatial average temperatures will be most sensitive to the proxy network during warm intervals. The relatively good agreement between different reconstructions during the MCA [*Jansen et al.*, 2007; *Masson-Delmotte et al.*, 2013] is likely a result of the short, common reference interval used to standardize the data in those studies [*Tingley*, 2012].

Correlations between the spatial mean and standard deviation are significant with respect to AR(1) null distributions that assume normality but do not appear significant against a rotational null hypothesis that preserve the time series properties of individual records while eliminating spatial correlations between them. A notable feature of the rotational null distributions for testing the significance of the observed correlations between means and standard deviations is that they are centered on positive values. One possible explanation is that the underlying distribution at each location features positive skew and is therefore not normal. Indeed, positively skewed temperature distributions naturally lead to larger standard deviations for samples that feature larger means [*Huybers et al.*, 2014], and we find that both the proxy and instrumental data set feature positive skew. As might be expected, temperature variability involves many moments, and here we have added to the discussion by formally investigating the second spatial moment.

Our results point to the need for further investigations of the distributional and space-time covariance properties of surface temperatures and surface temperature anomalies. If the past is to serve as a meaningful prologue for the future, the nonnormalities identified in surface temperature anomalies, as well as the possible role of nonstationary spatial covariance, must be better understood. It is likewise necessary to achieve a better understanding of the effects of changes in natural and anthropogenic forcings on the location, scale, shape, and space-time covariance structure of temperature anomaly distributions. Furthermore, it is important to delineate the internal and external contributions to the covariability of the mean and standard deviation and to determine if this relationship holds for intervals in the more distant past. Further work is also required to determine if the correlation between the mean and variability is a global feature, resulting from a common inherent structure of the temperature distribution, or if regional-scale phenomena, such as asymmetric Arctic amplification, are driving the observed behavior. The Coupled Model Intercomparison Project Phase 5 multimodel ensemble may prove useful in investigating these questions, as the covariability between the mean and standard deviation could be separately explored using historical simulations and control runs. Also important in this respect would be to better quantify the fidelity with which climate simulations are consistent with the higher moments found in observational data.

With the caveat that both the statistical description and physical understanding of surface temperature anomalies and their temporal evolution remain incomplete, we speculate that spatial variability will increase in tandem with future increases in annually averaged surface temperatures. Increased spatial variability would have important regional consequences because of the compound effects of a greater mean and greater variability in realizing high-temperature extremes.

Appendix A: Assumptions and Analysis Methods

A1. Probability Model

We assume there exists an unobserved, or latent, time series of Northern Hemisphere annual mean surface temperature anomalies, referred to as the *signal*, S_t. At each time, t, the spatial vector of N true, but latent, temperature anomaly values, $\mathbf{T_t}$, takes the form

$$\mathbf{T_t} = S_t \cdot \mathbf{1} + \boldsymbol{\epsilon}_\mathbf{t}, \tag{A1}$$

where S_t is the scalar signal at time t and $\mathbf{1}$ is an N vector of ones. The *spatial deviations* $\boldsymbol{\epsilon}_\mathbf{t}$ are mean-zero random vectors that covary in space and time, reducing the equivalent number of spatially independent samples, N_{eq}, at each time point, and the number of temporally independent samples at each location. Furthermore, we assume that the variance of the spatial deviations is dependent on time and possibly space as well; that is, $Var[\epsilon(s_i, t_j)] = \sigma^2_\epsilon(s_i, t_j)$ is a function of time and (possibly) space. We assume that the instrumental records can be represented as the true temperature anomalies with additive noise, and that the proxy records can be represented as linear transformations of the true temperature anomalies, with additive noise [cf. *Tingley and Huybers*, 2010a]. That is, for the instrumental data set we assume that

$$\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{t}} = S_t \cdot \mathbf{1} + \boldsymbol{\epsilon}_\mathbf{t} + \mathbf{e}^{\mathbf{I}}_{\mathbf{t}}, \tag{A2}$$

where $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{t}}$ is the vector of instrumental observations at time t and $\mathbf{e_t}$ is the corresponding vector of noise terms at time t. The formalism is the same for the proxies, but with an unknown linear transformation applied to $S_t \cdot \mathbf{1} + \boldsymbol{\epsilon}_\mathbf{t}$. The noise terms for the instruments are assumed to be independent in space and time so that $\text{Corr}[e^I(s_i, t_j), e^I(s_k, t_l)] = 0$ unless $i = k$ and $j = l$, and likewise for the proxy noise terms. As separate inference on the noise ($\mathbf{e_t}$) and spatial deviations ($\boldsymbol{\epsilon}_\mathbf{t}$) is not possible, estimates of the spatial standard deviation for each year are biased high by the presence of noise. As only changes in the standard deviation, rather than the absolute values, are of interest, the possible bias is only a concern insomuch as the additional noise term will widen confidence intervals. We assume that the noise variance for both the proxies and the instruments is constant through time, or at least has significantly smaller temporal variability than the spatial deviations. That is, we assume to first order that $\text{Var}[e^I(s_i, t_j)] = \sigma^2_e(s_i)$ does not depend on time and attribute any structure in the estimated spatial standard deviation time series to changes in the variance of the spatial deviations, $\sigma^2_\epsilon(s_i, t_j)$.

A2. Bootstrap Procedures and the Equivalent Number of Independent Samples

Confidence intervals are based on a modification of the bias corrected-accelerated (BCa) bootstrap procedure of *Efron and Tibshirani* [1986], using a minimum of 2000 bootstrap resamples. Spatial correlations, such as those we assume for the ϵ_t, result in standard bootstrap confidence intervals that are too narrow and that feature actual coverage rates that are lower than nominal. Experiments with surrogate data show that confidence intervals for estimates of the mean are correct provided that the equivalent number of spatially independent data points, N_{eq}, are resampled from the larger available pool of N data points at each bootstrap iteration and that N is sufficiently large (Figures A1 and A2). Given the correlation structure of a length N random vector, the equivalent number of independent samples, N_{eq}, can be calculated, following [*Liu*, 2001], as follows:

$$N_{\text{eq}} = \frac{N^2}{\sum_{i,j=1}^{N} \rho(x_i, x_j)}, \tag{A3}$$

where $\rho(x_i, x_j)$ is the correlation between data points x_i and x_j.

To explore the effects of spatial correlation on the BCa bootstrap confidence intervals, we make use of an experiment using simulated AR(1) series. If the length of an AR(1) series is N, and the number of independent samples is N_{eq}, then by plugging $\rho(x_i, x_j) = \alpha^{|i-j|}$ into equation (A3) it is possible to solve for the value of the AR(1) coefficient α as the positive, real root of,

$$N_{\text{eq}} \cdot \sum_{k=1}^{N-1} 2\,(N-k)\,\alpha^k + N\left(N_{\text{eq}} - N\right) = 0. \tag{A4}$$

To show the dependency of confidence interval coverage rates on the true value of N_{eq}, when resampling (with replacement) a fixed number N_{rs} of observations at each iteration of the bootstrap, we use the entire

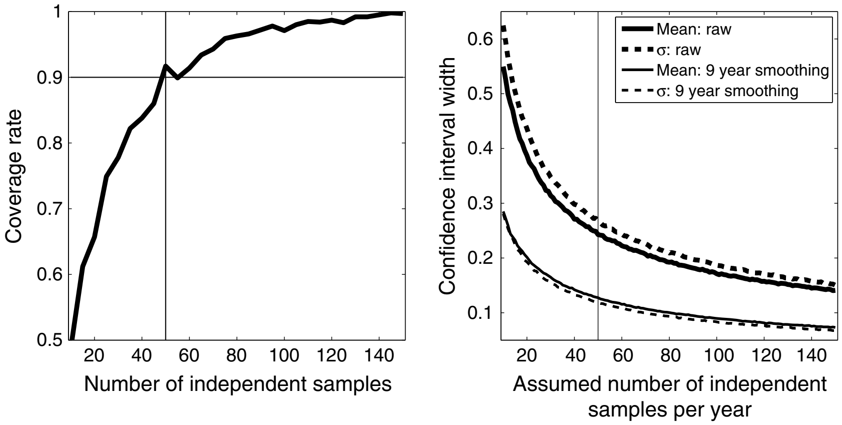

Figure A1. (left) Coverage rates of 90% BCa bootstrap confidence intervals for estimating the mean of surrogate data resembling the instrumental data, as a function of N_{eq}, the actual number of independent data points. $N_{rs} = 50$ of 624 data points are resampled at each iteration of the bootstrap. (right) Average 90% BCa bootstrap confidence interval widths for the mean and standard deviation of the instrumental data, using different values of N_{rs}, the number of observations resampled at each iteration of the bootstrap. There are a total of 624 observations each year. Straight and dashed lines correspond to the mean and standard deviation, respectively; thick and thins lines correspond to results for the raw and smoothed series, respectively.

instrumental data set to arrive at a pooled estimate of the spatial standard deviation and then construct 1000 AR(1) data sets of length 624 (the number of instrumental locations) with this variance for a given value of N_{eq}. That is, we simulate AR(1) time series with length given by the spatial dimension of the instrumental data. We vary the equivalent number of independent samples between 10 and 150, corresponding to α values between 0.97 and 0.61. For each value of N_{eq} we find 90% BCa bootstrap confidence intervals for the mean by repeatedly resampling (with replacement) $N_{rs} = 50$ of the 624 values. Calculating the proportion of the 1000 confidence intervals that cover the true mean value (Figure A1) shows that the coverage rate is close to the nominal value of 0.9 when $N_{eq} = N_{rs} = 50$. When $N_{eq} < N_{rs}$, the coverage rate is lower than 0.9 (confidence intervals too narrow), while when $N_{eq} > N_{rs}$, the coverage rate is larger than 0.9 (confidence intervals too wide).

As N_{eq} for the instrumental data is unknown, we explore the dependency of confidence interval widths on the value of N_{rs}, the number of data points resampled from the original 624 spatial observations at each iteration of the bootstrap. The confidence intervals narrow with increasing values of N_{rs}, and with increasing temporal

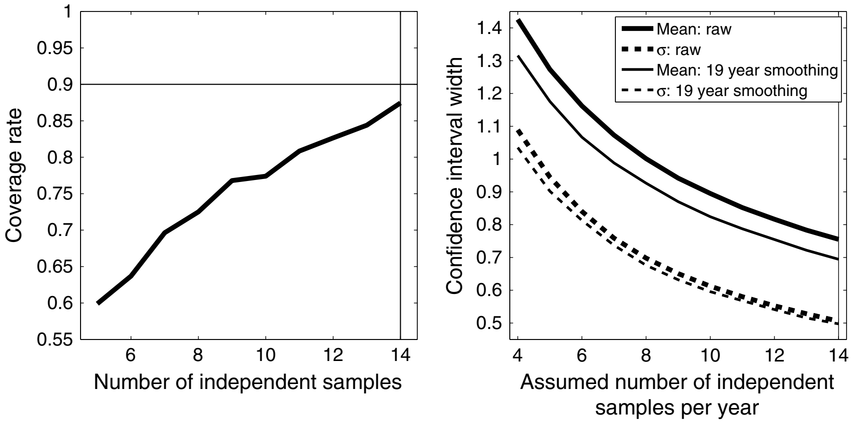

Figure A2. (left) Coverage rates of 90% BCa bootstrap confidence intervals for estimating the mean of surrogate data resembling the proxy data, as a function of N_{eq}, the actual number of independent data points. $N_{rs} = 14$ of 14 data points are resampled at each iteration of the bootstrap. (right) Average (over the 1601–1800 interval) 90% BCa bootstrap confidence interval widths for the mean and standard deviation of the proxy data, assuming different values of N_{rs}, the number of observations resampled at each iteration of the bootstrap. There are a total of 14 records each year. Straight and dashed lines correspond to the mean and standard deviation, respectively; thick and thin lines correspond to results for the raw and smoothed series, respectively.

smoothing (Figure A1). For the analysis reported above, we resample $N_{rs} = 50$ of the available 624 observations: we assume $N_{eq} = 50$. Reducing the value of N_{rs} to 25 does not qualitatively affect our conclusions about the relative warmth and variability of different intervals in the twentieth century, as intervals can still be readily distinguished on the basis of the mean, and even using $N_{rs} = 50$ we have not argued that the different periods can be distinguished on the basis of the standard deviation (Figure 3). A value between 25 and 50 for the equivalent number of spatially independent data points in the Northern Hemisphere is in line with other estimates for similarly sized regions of the Earth [*Stine et al.*, 2008; *Tingley and Huybers*, 2013].

Similar simulations mimicking the proxy data (Figure A2) show that coverage rates for 90% confidence interval for the mean are biased marginally low, even when setting $N_{eq} = N_{rs} = 14$—a result of the small sample size and the bootstrap-derived confidence intervals being exact only in the large data limit. For the proxy analysis, we set N_{rs} equal to the number of observations available at each year (maximum of 14). From the probability model, the common dependency on S_t should result in higher correlations between the data records than between the time series of spatial deviations—which determines the number of spatially independent samples. The correlations between the proxy records themselves are, however, very low—only one pairwise correlation is larger than 0.35, nine are larger than 0.2, and 21 are less than zero—justifying our decision to resample the number of available observations at each year, with replacement, when performing the bootstraps. We judge the error introduced to the coverage probabilities by the small sample size to be less important than the fact that the BCa bootstrap is robust to departures from normality. As the proxy records are restandardized after smoothing, the confidence intervals for the mean and standard deviation do not display the dramatic narrowing with increasing smoothing seen in analysis of the instrumental data (compare Figure A1 and Figure A2).

A2.1. Confidence Intervals for Correlations

To estimate uncertainty in correlation coefficients, we use the BCa bootstrap, resampling $2N/L$ of the original mean and standard deviations pairs, where N is the length of the series and L the length of the smoothing window. The rationale behind this choice is twofold. First, the smoothing window reduces the number of temporally independent points by a factor less than the window length. Second, the time series of standard deviations across the smoothed records retains higher frequencies than the corresponding mean time series. Indeed, experiments based on smoothing independent normal deviates show that the spectrum of the standard deviation time series retains more power at higher frequencies than the mean time series. As a result, the standard deviation time series has more independent samples than the mean time series, which, in turn, has more independent samples than N/L. Furthermore, equating the 95% upper critical value under the AR(1) null to the standard expression for this critical value [*Zar*, 1999] and solving for the sample size gives a similar result to the $2N/L$ rule (Table 2). We make the same assumption when finding the uncertainties in means and standard deviations averaged over different intervals (Figure 6).

Acknowledgments
Proxy data are available through the NOAA Paleoclimatology portal at www.ncdc.noaa.gov/paleo/pubs/moberg2005/moberg2005.html [*Moberg et al.*, 2005] and www.ncdc.noaa.gov/paleo/pubs/osborn2006/osborn2006.html [*Osborn and Briffa*, 2006]. The instrumental compilation is available at www.cru.uea.ac.uk/cru/data/temperature/. This work is supported in part by NSF P2C2 grant 1304309.

References

Bradley, R. S., M. K. Hughes, and H. F. Diaz (2003), Climate in Medieval time, *Science*, *302*, 404–405.

Briffa, K. R., F. H. Schweingruber, P. D. Jones, T. J. Osborn, S. G. Shiyatov, and E. A. Vaganov (1998), Reduced sensitivity of recent tree-growth to temperature at Northern high latitudes, *Nature*, *391*, 678–682.

Broeker, W. S. (2001), Was the Medieval warm period global?, *Science*, *291*, 1497–1499.

Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *J. Geophys. Res.*, *111*, D12106, doi:10.1029/2005JD006548. [Available at http://www.cru.uea.ac.uk/cru/data/temperature/.]

Christiansen, B. (2011), Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?, *J. Clim.*, *24*, 674–692.

Collins, M., et al. (2013), Long-term climate change: Projections, commitments and irreversibility, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. Stocker et al., pp. 1029–1136, Cambridge Univ. Press, Cambridge, U. K., and New York., doi:10.1017/CBO9781107415324.024.

Cressie, N., and M. P. Tingley (2010), Discussion of "The value of multi-proxy reconstruction of past climate" by B. Li, D.W. Nychka, and C.M. Ammann, *J. Am. Stat. Assoc.*, *105*(491), 895–900.

Cronin, T., G. Dwyer, T. Kamiya, S. Schwede, and D. Willard (2003), Medieval warm period, Little Ice Age and 20th century temperature variability from Chesapeake Bay, *Global Planet. Change*, *36*, 17–29.

Crowley, T., and T. Lowery (2000), How warm was the Medieval Warm Period?, *Ambio*, *29*, 51–54.

D'Arrigo, R. D., G. C. Jacoby, D. Frank, N. Perderson, E. Cook, B. Buckley, B. Nachin, R. Mijiddorf, and C. Dugarjav (2001), 1738 years of Mongolian temperature variability inferred from a tree-ring chronology of Siberian pine, *Geophys. Res. Lett.*, *28*(3), 543–546.

D'Arrigo, R. D., R. Wilson, and G. C. Jacoby (2006), On the long-term context for late twentieth century warming, *J. Geophys. Res.*, *111*, D03103, doi:10.1029/2005JD006352.

Diaz, H. F., R. Trigo, M. K. Hughes, M. E. Mann, E. Xoplaki, and D. Barriopedro (2011), Spatial and temporal characteristics of climate in Medieval times revisited, *Bull. Am. Meteorol. Soc.*, *92*(11), 1487–1500.

Efron, B., and R. Tibshirani (1986), Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Stat. Sci.*, *1*(1), 54–75.

Esper, J., and D. Frank (2009), The IPCC on a heterogeneous medieval warm period, *Clim. Change*, *94*(3–4), 267–273.

Esper, J., E. Cook, and F. Schweingruber (2002), Low frequency signals in long tree-ring chronologies for reconstructing past temperature variability, *Science*, *295*, 2250–2253.

Goose, H., H. Renssen, A. Timmermann, and R. S. Bradley (2005), Internal and forced climate variability during the last millennium: A model-data comparison using ensemble simulations, *Quat. Sci. Rev.*, *24*, 1345–1360.

Goose, H., O. Arzel, M. E. Mann, H. Renssen, N. Riedwyl, A. Timmermann, E. Xoplaki, and H. Wanner (2006), The origin of the European "Medieval Warm Period", *Clim. Past*, *2*, 99–113.

Guiot, J., C. Corona, and ESCARSEL members (2010), Growing season temperatures in Europe and climate forcings over the past 1400 years, *PloS one*, *5*(4), e9972.

Hansen, J., M. Sato, and R. Ruedy (2012), Perception of climate change, *Proc. Natl. Acad. Sci.*, *109*, E2415–E2423.

Hartmann, D., et al. (2013), Observations: Atmosphere and surface, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. Stocker et al., pp. 159–254, Cambridge Univ. Press, Cambridge, U. K, and New York., doi:10.1017/CBO9781107415324.008.

Huang, S., H. N. Pollack, and P. Y. Shen (1997), Late Quaternary temperature changes seen in world-wide continental heat flow measurements, *Geophys. Res. Lett.*, *24*(15), 1947–1950.

Hughes, M. K., and H. F. Diaz (1994), Was there a "Medieval Warm Period", and if so, where and when?, *Clim. Change*, *26*(2–3), 109–142.

Huntingford, C., P. D. Jones, V. N. Livina, T. M. Lenton, and P. M. Cox (2013), No increase in global temperature variability despite changing regional patterns, *Nature*, *500*, 327–330.

Huybers, P., A. Rhines, K. McKinnon, and M. P. Tingley (2014), U.S. daily temperatures: The meaning of extremes in the context of non-normality, *J. Clim.*, *27*(9), 7368–7384.

Intergovernmental Panel on Climate Change (2007), Summary for policymakers, in *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*, edited by C. Field et al., Cambridge Univ. Press, Cambridge, U. K., and New York.

Jacoby, G., and R. D'Arrigo (1995), Tree ring width and density evidence of climatic and potential forest change in Alaska, *Global Biogeochem. Cycles*, *9*(2), 227–234.

Jansen, E., et al. (2007), Palaeoclimate, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., chap. 6, pp. 435–484, Cambridge Univ. Press, Cambridge, U. K., and New York.

Katz, R. W., and B. G. Brown (1992), Extreme events in a changing climate: Variability is more important than averages, *Clim. Change*, *21*(3), 289–302.

Lamb, H. H. (1965), The early Medieval Warm Epoch and its sequel, *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, *1*(13), 13–37.

Li, B., D. Nychka, and C. Ammann (2010), The value of multi-proxy reconstruction of past climate, *J. Am. Stat. Assoc.*, *105*(491), 883–911. With discussions and rejoinder.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, Springer, New York.

Ljungqvist, F. C. (2010), A new reconstruction of temperature variability in the extra-tropical Northern Hemisphere during the last two millennia, *Geogr. Ann., Ser. A*, *92*(3), 339–351.

Mann, M., Z. Zhang, S. Rutherford, R. Bradley, M. Hughes, D. Shindell, C. Ammann, G. Faluvegi, and F. Ni (2009), Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, *Science*, *326*(5957), 1256–1260.

Masson-Delmotte, V., et al. (2013), Information from paleoclimate archives, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. Stocker et al., Cambridge Univ. Press, Cambridge, U. K., and New York.

Moberg, A., D. Sonechkin, K. Holmgren, N. Datsenko, and W. Karlén (2005), Highly variable northern hemisphere temperatures reconstructed from low-and high-resolution proxy data, *Nature*, *433*(7026), 613–617.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *J. Geophys. Res.*, *117*, D08101, doi:10.1029/2011JD017187.

Osborn, T. J., and K. R. Briffa (2006), The spatial extent of 20th-century warmth in the context of the past 1200 years, *Science*, *311*, 841–844.

Rahmstorf, S., and D. Coumou (2011), Increase of extreme events in a warming world, *Proc. Natl. Acad. Sci.*, *108*(44), 17,905–17,909.

Rhines, A., and P. Huybers (2013), Frequent summer temperature extremes reflect changes in the mean, not the variance, *Proc. Natl. Acad. Sci.*, *110*(7), E546.

Stine, A., and P. Huybers (2014), Arctic tree rings as recorders of variations in light availability, *Nat. Commun.*, *5*, 3836.

Stine, A., P. Huybers, and I. Fung (2008), Changes in the phase of the annual cycle of surface temperature, *Nature*, *457*(7228), 435–440.

Tingley, M. (2012), A Bayesian ANOVA scheme for calculating climate anomalies, with applications to the instrumental temperature record, *J. Clim.*, *25*, 777–791. [Available at ftp://ftp.ncdc.noaa.gov/pub/data/paleo/ softlib/anova.]

Tingley, M., and P. Huybers (2010a), A Bayesian algorithm for reconstructing climate anomalies in space and time: Part 1. Development and applications to paleoclimate reconstruction problems., *J. Clim.*, *23*(10), 2759–2781.

Tingley, M., and P. Huybers (2010b), A Bayesian algorithm for reconstructing climate anomalies in space and time: Part 2. Comparison with the regularized expectation-maximization algorithm, *J. Clim.*, *23*(10), 2782–2800.

Tingley, M., and B. Li (2012), Comments on "Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?", *J. Clim.*, *25*(9), 3441–3446.

Tingley, M., P. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam (2012), Piecing together the past: Statistical insights into paleoclimatic reconstructions, *Quat. Sci. Rev.*, *35*, 1–22.

Tingley, M. P., and P. Huybers (2013), Recent temperature extremes at high northern latitudes unprecedented in the past 600 years, *Nature*, *496*(7444), 201–205.

Tingley, M. P., A. R. Stine, and P. Huybers (2014), Temperature reconstructions from tree-ring densities overestimate volcanic cooling, *Geophys. Res. Lett.*, *41*, 7838–7845, doi:10.1002/2014GL061268.

van Engelen, A. F. V., J. Buisman, and F. IJnsen (2001), A millennium of weather, winds and water in the low countries, in *History and Climate: Memories of the Future?*, edited by P. D. Jones et al., pp. 101–124, Kluwer Academia, Norwell, Mass.

Yang, B., A. Braeuning, K. R. Johnson, and S. Yafeng (2002), General characteristics of temperature variation in China during the last two millennia, *Geophys. Res. Lett.*, *29*(9), 1324, doi:10.1029/2001GL014485.

Zar, J. H. (1999), *Biostatistical Analysis*, 4th ed., Pearson Edu., Singapore.