RECONSTRUCTING PAST CLIMATE FROM NATURAL PROXIES AND ESTIMATED CLIMATE FORCINGS USING SHORT AND LONG-MEMORY MODELS

BY LUIS BARBOZA*, BO LI†, MARTIN P. TINGLEY‡ AND FREDERI G. VIENS§

Universidad de Costa Rica*, University of Illinois at Urbana-Champaign†, Pennsylvania State University‡ and Purdue University§

We produce new reconstructions of Northern Hemisphere annually averaged temperature anomalies back to 1000AD, based on a model that includes external climate forcings and accounts for any long-memory features. Our reconstruction is based on two linear models, with the first linking the latent temperature series to three main external forcings (solar irradiance, greenhouse gas concentration, and volcanism), and the second linking the observed temperature proxy data (tree rings, sediment record, ice cores, etc.) to the unobserved temperature series. Uncertainty is captured with additive noise, and a rigorous statistical investigation of the correlation structure in the regression errors motivates the use of long memory fractional Gaussian noise models for the error terms. We use Bayesian estimation to fit the model parameters and to perform separate reconstructions of land-only and combined land-and-marine temperature anomalies. We quantify the effects of including the forcings and long memory models on the quality of model fits, and find that long memory models can result in more precise uncertainty quantification, while the external climate forcings substantially reduce the squared bias and variance of the reconstructions. Finally, we use posterior samples of model parameters to arrive at an estimate of the transient climate response to greenhouse gas forcings of 2.56°C (95% credible interval of [2.20, 2.95]°C), in line with previous, climate-model-based estimates.

1. Introduction. An understanding of recently observed and projected future climate changes (Alexander et al., 2013) within the context of the natural variability and dynamics of the climate system requires accurate and precise reconstructions of past climate. As spatially wide-spread instrumental temperature observations extend back to only about 1850, it is necessary to turn to the noisy and sparsely distributed paleoclimate record to characterize natural climate variability on longer time scales. While there is

AMS 2000 subject classifications: Primary 62P12; secondary 60G22

Keywords and phrases: external forcings, long-memory, proxies, temperature reconstruction
now a rich tradition of inferring past climate from natural proxies, such as tree rings, corals, ice cores, lake floor sediment cores, and measurement on speleothems (for recent reviews, see NRC, 2006; Jones et al. 2009; Tingley et al. 2012), many scientific and statistical challenges remain.

1.1. Paleoclimatology context. Reconstructions of past surface temperatures from networks of multiple proxy types are prevalent in the climate science literature of the last 15 years – notable examples include Overpeck et al. (1997), Mann, Bradley and Hughes (1998, 1999), Luterbacher et al. (2004), Moberg et al. (2005), Juckes et al. (2006), Mann et al. (2008a, 2009), Kaufman et al. (2009), Tingley and Huybers (2013), and PAGES 2k Consortium (2013). While these studies have substantially increased our understanding of past climate, limitations remain in terms of the statistical treatment and uncertainty quantification. As described in Tingley et al. (2012), the most commonly used approaches to paleoclimate reconstruction are all variants of multiple linear regression (see, for example, Table 1 of Christiansen, Schmith and Thejll, 2009), regularized in some fashion to account for the “p > n” problem in the estimation procedure. Examples of particularly popular estimation approaches include regularized variants of the Expectation-Maximization algorithm (Dempster, Laird and Rubin, 1977; Schneider, 2001; Rutherford et al., 2003, 2005; Zhang, Mann and Cook, 2004; Mann et al., 2007, 2005; Steig et al., 2009), and principal component regression (Cook, Briffa and Jones, 1994; Mann, Bradley and Hughes, 1998; Luterbacher et al., 2004; Wahl and Smerdon, 2012), which is sometimes combined with canonical correlation analysis (Smerdon et al., 2010). A common shortcoming of these studies lies in the limited propagation of parameter uncertainty through the model, including uncertainty in the estimation of regularization parameters; for further discussion see Schneider (2001), Smerdon et al. (2010), and the supplement to Wahl and Smerdon (2012).

Recently, hierarchical modeling and Bayesian inference techniques have been proposed and employed to reconstruct past climate from proxies (Haslett et al., 2006; Li, Nychka and Ammann, 2010; Tingley and Huybers, 2010a,b; Werner, Luterbacher and Smerdon, 2013; Tingley and Huybers, 2013). Hierarchical modeling is a natural framework for including the available scientific understanding of both the target climate process (e.g., annual surface temperature anomalies), and how the various natural proxies are causally affected by variations in the climate system. Bayesian inference, in turn, provides a cohesive framework for propagating uncertainty, while the posterior draws of the target climate quantity are a more statistically precise and
scientifically useful result than a point estimate and associated uncertainty interval (Tingley et al., 2012).

In this paper, we reconstruct Northern Hemisphere (NH) temperature anomalies over the past millennium using a hierarchical Bayesian model that describes temperature as linearly dependent on three important climate forcings: greenhouse gas concentrations, volcanic aerosol concentrations, and variations in solar irradiance. The proxies, in turn, are modeled as linear in the latent temperature process. At both levels, we model the errors using long-memory processes due to the statistical evidence of long-range correlation exhibited in the errors. To our knowledge, this is the first ensemble-based paleoclimate reconstruction that includes the effects of climate forcings, and the first application of long-memory noise processes to the temperature reconstruction. As our method involves first reducing the proxy data set to a single time series, and then inferring hemispheric average temperature anomalies, rather than the spatial pattern, our analysis is a form of composite-plus-scaling (Tingley et al., 2012).

The external forcings used in the analysis are closely related to global temperature evolution. The Intergovernmental Panel on Climate Change (IPCC) has steadily increased its certainty level on stating the causal relationship between increasing atmospheric concentrations of anthropogenic greenhouse gases and increasing average global temperatures, reaching the “extremely likely” level of 95% confidence in 2013 (AchutaRao et al., 2013). The relationship between solar irradiance and surface temperatures is studied in Crowley and Kim (1996), Lean, Beer and Bradley (1995), while Briffa et al. (1998), Crowley and Kim (1993), Crowley, Criste and Smith (1993) and Landrum et al. (2013) analyzed the effect of volcanic activity on global temperatures. Furthermore, in a conceptual study using climate model data, Li, Nychka and Ammann (2010) demonstrated that temperature reconstructions are improved when information about the climate forcing is included in the reconstruction. We therefore include all three major external forcings in our reconstructions.

1.2. Long-memory modeling and estimation challenges. To our knowledge, the error terms in all previous models for multi-proxy climate reconstructions are assumed to be white or autoregressive (AR; see, for example, Tingley et al., 2012). For instance, Li, Nychka and Ammann (2010), Tingley and Huybers (2010a) and McShane and Wyner (2011) use AR(1) or AR(2) errors, while reconstructions based on the expectation-maximization algorithm or principal component regression have generally not explicitly modeled temporal autocorrelation (see section 8.7.4 of Tingley et al., 2012).
While graphs of the autocorrelation and partial autocorrelation functions (acf and pacf) are helpful in choosing the structure of the error model, and are adequate for identifying long memory in simple models, they can be inadequate for assessing long memory behavior in hierarchical models. In such cases, short data streams disallow reliance on known asymptotic properties, and lack of self-similarity means that inference on one range of frequencies cannot apply to another. These issues are well known for widely used long-memory time series models, such as fractional autoregressive integrated moving average (FARIMA) models (Beran, 1994). Misspecification of a long memory process with a short memory model can lead to erroneously attributing long-memory effects to deterministic trends or external forcings, and thus will affect uncertainty quantification. Specifically, since long-memory models can exhibit larger asymptotic variances than their relatively short memory model analogues (see Chronopoulou, Viens and Tudor, 2009, and references therein), reported uncertainty levels under memory misspecification can be lower than the nominal values.

To formally explore evidence for long memory in both the proxy data and the modern temperature record, we use statistical techniques specifically developed for detecting long-range correlations.

Motivated by the limitations of the data, and our goal of using a robust model, we focus on the simplest long-memory model available in our setting: linear regression with fractional Gaussian noise (fGn) errors. While the fGn errors are self-similar, the resulting hierarchical linear model is not. The theoretical question of estimating memory length for non-self-similar models is notoriously difficult. Asymptotic theory is still under development, and current work on high-frequency or increasing-horizon versions of our model can not yet be considered definitive. Section ?? provides brief background information on long-memory estimation, while further details can be found in references therein; see in particular Gneiting and Schlather (2004).

In the context of annual paleoclimate observations, time intervals cannot be assumed small, and the calibration period is short. On account of the long time intervals, we cannot use the local path behavior of the data (e.g. Hölder continuity) as a proxy for long memory – an approach that is possible for fGn-driven models where high frequency data exists. Such models are asymptotically Hölder-continuous in the limit of ultra-high frequency, with a parameter that also governs long memory. On account of the short calibration period, methodologically sound results from low frequency increasing-horizon asymptotics (see Tudor and Viens, 2007) cannot be used to measure long-range dependence in our case, as there is simply not enough data. Instead we resort to a fully Bayesian framework to estimate all param-
etters, including those responsible for memory length, with the added benefit of a complete evaluation and propagation of uncertainty.

This article is structured as follows. Section 2 describes the datasets used in the reconstruction, and Section 3 gives the details of the hierarchical Bayesian models. Section 4 presents the results of our Bayesian reconstructions. We compare our results with previous reconstructions and discuss the estimation of transient climate response in Section 5 before summarizing our quantitative conclusions and discussing remaining challenges in Section 6. Two Online Supplements provide further details on the modeling framework and additional quantitative results.

2. Data sets. The analysis makes use of three distinct data sources: observed temperature anomalies (in °Celsius) over the period 1900-1998; a suite of temperature-sensitive proxies over the period 1000-1998 taken from the database originally described in Mann et al. (2008a) and used additionally in Mann et al. (2009); and estimates of external climate forcings from 1000-1998 AD.

We make use of two different instrumental estimates of post-1850 NH temperature anomalies, both developed by the Climate Research Unit of the University of East Anglia (Brohan et al., 2006). The CRUTEM3v data set (abbreviated hereafter as CRU) is an estimate of air surface temperature anomalies over land, while HadCRUT3v (hereafter abbreviated as HAD) is an estimate of combined land air- and marine sea- surface temperatures. These data sets are widely used for the calibration of proxy-based climate reconstructions (e.g. Mann et al., 2008a; Luterbacher et al., 2004; Rutherford et al., 2005; Kaufman et al., 2009; McShane and Wyner, 2011; Tingley and Huybers, 2013). We make use of the variance-adjusted version of each data set to facilitate comparisons with results from Mann et al. (2008a). While both instrumental data sets extend back to 1850, we choose 1900-1998 as our calibration period, as the sparsity of instrumental observations results in less trustworthy Hemispheric estimates prior to about 1900 (Smith, 2010).

The proxies used in our analysis are selected from the 1,209 climate-sensitive proxies originally compiled in Mann et al. (2008a)\(^1\). This compilation brings together a wide array of proxy types, including tree ring widths and densities, marine sediment cores, speleothems (cave deposits), lacustrine sediment cores, ice cores, coral records, and historical documentary information (see NRC, 2006 and Jones et al., 2009, for further descriptions of

\(^1\)For more details on the dataset, see the NOAA-Paleoclimatology/World Data Center at: http://www.ncdc.noaa.gov/paleo/pubs/pcn/pcn-proxy.html.
each of these data types). The proxy data are not raw observations, but are rather processed to remove non-climatic variability, such as age effects associated with tree ring data. This type of processing results in a data product which may be more directly interpreted as “climate sensitive”, according to the paleoclimatology community. While it is common to base climate reconstructions on the post-processed data, as is done here, we acknowledge that doing so does neglect the uncertainty inherent in the processing steps. We set aside for future research the challenge of including the processing of raw climate proxy observations into climate-sensitive series, but note that such steps can theoretically be included within the hierarchical framework developed here. For further details concerning the processing of raw proxy observations see, for example, NRC (2006); Jones et al. (2009).

Estimates of the external climate forcings – atmospheric greenhouse gas concentrations (C), solar irradiance (S), and volcanism (V) – are described and plotted in Li, Nychka and Ammann (2010) and described more fully in Ammann et al. (2007). The original greenhouse gas concentration time series is in units of CO$_2$ equivalent in parts per million; the solar irradiance series is in Watt/m$^2$ and the estimated volcanic series is in units of teragrams of sulfate per year (see Ammann et al., 2007, for further details).

3. **Model specification.** Hierarchical Bayesian models typically consist of three levels. The data level describes the likelihood of the observations conditional on a latent stochastic process. In our context, the latent process is the time series of NH mean temperature anomalies, and the observations are the proxies. The process level describes the parametric structure of the latent process – often with recourse to prior scientific information, such as knowledge of the underlying physical dynamics (e.g. Berliner, Wikle and Cressie, 2000). Finally, the prior level provides closure and allows for Bayesian inference by providing prior distributions for all unknown parameters in the data- and process-levels. For a general description of hierarchical modeling and Bayesian inference in the paleoclimate context, see Tingley et al. (2012). Following Li, Nychka and Ammann (2010), the data level models the proxies as a normal distribution with mean equal to a linear function of the latent, unobserved true temperatures, while the process level models the latent temperature process as normal with mean given by a linear function of the external forcings (Li et al., 2010). We add to previous work by applying the model to actual proxy data, as opposed to using pseudo proxy experiments derived from climate model output (Li, Nychka and Ammann, 2010), as well as identifying and modeling long-memory error structures in both levels.
The Bayesian modeling framework is closely related to the stochastic filtering methodology. An interesting application of classical explicit Kalman filtering (see Kalman and Bucy, 1961) to climatic reconstruction is in Lee, Zwiers and Tsao (2008), where the authors use forcings and a smaller proxy dataset to reconstruct temperatures on a decadal basis. However, there are, to our knowledge, no practical tools for filtering with fGn errors, and in addition, stochastic filters, which are adapted to tracking moving signals dynamically in time, are notoriously poor at estimating fixed parameters; see Yang et al. (2008) and Chronopoulou and Viens (2012). Thus they are poor choices for our paleo long-memory setting, whereas the Bayesian approach adopted here allows for all parameters to be estimated simultaneously while avoiding the known estimation difficulties inherent to filtering. Moreover, since the proxy observations are not being updated over time, the sequential updating property of filtering is not advantageous.

3.1. Proxy data reduction. It is desirable for several reasons to reduce the dimensionality of the proxy data set, which consists of 1,209 time series. First, as there are only a limited number of years in the calibration interval, dimensionality reduction can lead to a more parsimonious model, avoid over fitting, and lead to more robust temperature reconstructions. Second, our interest in inferring global mean temperatures rather than spatial fields motivates a reduction, prior to fitting a hierarchical model, to a single time series that reflects the shared variability between the proxies that is likely attributable to a common, climatic origin. Third, the proxy reduction is important in limiting the computational burden of estimating parameters describing long-memory; for a comparison between computational and asymptotic efficiency for various long-memory parameter estimators, see Chronopoulou and Viens (2009). We therefore perform a sequence of procedures to reduce the number of proxies while attempting to retain the useful information to a large extent.

Following Mann et al. (2008a), we first select only those proxies that are recorded at least as far back as 1000 A.D. and in addition have a significant correlation with their closest instrumental time series (marine or land) over their period of mutual overlap 1896-1995. We use local temperature information in the screening procedure as any proxy that might correlate to hemispheric temperature with some degree of accuracy should relate to its local temperature with higher precision (Mann et al., 2008a). Such a criterion does not take into account the possibility of exploiting physical teleconnections that exist in the actual climate system (Mann, Bradley and Hughes, 1998; Tingley et al., 2012; Werner, Luterbacher and Smerdon, 2013).
Table 1

Geographical distribution of the 38 proxies by type.

| Type          | #  | Locations                                      |
|---------------|----|-----------------------------------------------|
| Tree Ring     | 16 | USA, Argentina, Norway, New Zealand, Poland, Sweden |
| Lacustrine    | 7  | Mexico, Ecuador, Finland                      |
| Speleothem    | 6  | China, Scotland, Yemen, Costa Rica, South Africa |
| Ice cores     | 4  | Peru, Greenland, Canada                       |
| Other*        | 5  | China, Mongolia, Tasmania, New Zealand         |

* The category named “Other” contains data from composite temperature reconstructions and historical documentary series.

Fig 1: Geographical distribution of the 25 proxy series.

This screening procedure yields 38 proxies whose distribution by type and location is given in Table 1. Tree rings represent the majority of proxies that pass the screening criteria, consistent with the ubiquitous use of tree ring information in annual resolution temperature reconstructions (NRC, 2006; Jones et al., 2009; PAGES 2k Consortium, 2013, and references therein).

A number of the 38 proxy series in Table 1 show undesirable properties given our assumption of a stationary relationship between the proxies and temperatures. In particular, several of the lacustrine and speleothem records feature much greater variability in the early portion of the time interval than in the calibration period. On such bases, we exclude 13 proxies, leaving a total of 25; see Figure ?? and Table ?? in the Online Supplement ?? for details. The single lacustrine proxy included in the reconstructions is the tiljander 2003 darksum series from Finland (Tiljander et al., 2003). We apply a log-transformation on this series in order to dampen the few years that feature very thick varves (Loso, 2009), and to produce a series that is in-line with the assumption of normal errors in our statistical models. Figure 1 shows the spatial locations of the 25 proxies.

To increase computational tractability, and to ensure that the heteroge-
nous spatial distribution of the proxies does not bias estimates of the spatial average, we further reduce the 25 proxies into a single series, termed the “reduced proxy,” via a weighted averaging procedure. Intuitively, we seek a reduced proxy series that captures the common signal of globally averaged climate reflected in the shared variability between the proxies. We estimate the averaging weights used to form the reduced proxy using least squares regression, first centering and scaling each of the 25 proxy series over the period 1000-1998. Denoting these scaled proxies as $P_{i,t}, i = 1, \ldots, 25$ and $t = 1000, \ldots, 1998$ and the HAD or CRU series as $T_t$ (mean temperature anomalies), we estimate the weights via an ordinary least squares fit to $T_t = a_0 + \sum_{i=1}^{25} a_i P_{i,t} + \epsilon_t$, where $\epsilon_t$ is white noise. Since most of the proxies end after 1982, here we fit the model using only the data from 1900 to 1982. The least squares parameter estimates $\hat{a}_0, \ldots, \hat{a}_{25}$ provide a weighted average of proxies that maximizes the explained variance. Denote the reduced proxy as $RP_t$, then

$RP_t = \hat{a}_0 + \sum_{i=1}^{25} \hat{a}_i P_{i,t}$.

The percentage of variation in temperatures that can be explained by the reduced proxy is $R^2 = 77.48\%$ for the HAD data set and $R^2 = 58.25\%$ for the CRU data set; note that the $R^2$ is higher for the HAD data set despite all proxies being terrestrial. Some, but by no means all, of the proxies are coastal, which affords a partial explanation. The proxies are selected on the basis of local correlations, and the higher percentage of explained variation with the HAD data set is indicative of the fact that temperature observations at the locations of the proxies are better at predicting global land and sea temperature than global land-only temperatures. Note that colinearity will not be an issue here because on the one hand there is no strong correlation between $P_{i,t}$ and on the other hand we are mainly interested in the linear combination of $P_{i,t}$ rather than the coefficients $\hat{a}_i$.

The geophysical distribution of the weights (in percentage of absolute value) are displayed in Tables ?? and ?? in the Online Supplement ?? For both HAD and CRU data sets, proxies in the United States are most heavily weighted, followed by the Mongolian composite. The remaining countries have a fairly uniform distribution with no single country exceeding the 8% level (HAD) or 7% level (CRU). Our selected proxies therefore have broad spatial coverage, inasmuch as possible with the available data. The weights heavily concentrate on the “Tree Rings” and “Other” categories, consistent once more with the prevalence of tree ring series in climate reconstructions (e.g., Overpeck et al., 1997; Mann, Bradley and Hughes, 1998; Luterbacher
et al., 2004; Moberg et al., 2005; Tingley and Huybers, 2013; PAGES 2k Consortium, 2013). The weight for the single lacustine series, from Tiljander et al. (2003) is less than 8% for both HAD and CRU data sets, indicating that it exerts a limited control on the overall reconstructions. The limited influence of this lacustrine series is of particular importance given the known difficulties in its calibrating, due to the potential of anthropogenic impact on the lake catchment (Tiljander et al., 2003; Mann et al., 2008b); we return to this point in Section 4.2.

The modeling approach taken here, based on a weighted average of proxies that pass a local screening condition, does not explicitly consider long-range spatial dependencies, or teleconnections, within the climate system. Another option would be to set the reduced proxy to the leading principal component of the 25 proxies that pass the screening test. Such an approach would extract the dominant common signal shared by the proxies, whereas for the purposes of this analysis we are more interested in retaining the common temperature signal they share. While methods based on principal component or canonical correlation analysis are prevalent in paleoclimatology, both for the reconstruction of spatial patterns and (as here) spatial averages, there is ongoing debate as to the merits of such methods; see Cook, Briffa and Jones (1994); NRC (2006); Wahl and Smerdon (2012); Tingley et al. (2012); Werner, Luterbacher and Smerdon (2013); PAGES 2k Consortium (2013) for discussion.

### 3.2. Possible long memory behavior in the proxy data.

While the temperature-proxy relationship is almost universally assumed to be linear (e.g., Luterbacher et al. (2004), Rutherford et al. (2005), Li, Nychka and Ammann (2010), Tingley and Huybers (2010b), Kaufman et al. (2009), McShane and Wyner (2011), Christiansen (2011), Smerdon et al. (2010), and each of the methods in Table 1 of Christiansen, Schmith and Thejll (2009) and discussed in Section 8 of Tingley et al. (2012)), the correlation structure in the error term has not been thoroughly studied. The choice of model for the correlation structure is of particular importance as its adequacy directly affects the accuracy and precision of the uncertainty quantification associated with the reconstruction. Here we consider models of the form,

\[
RP_t = \alpha_0 + \alpha_1 T_t + \sigma_p \eta_t,
\]

where \( \eta_t \) is a zero-mean, unit-variance stationary stochastic process, and \( \sigma_p \) a constant variance parameter. We fit model (3.2) using least-squares over the 1900–1982 interval, using either the HAD or CRU as \( T_t \), and examine the correlation structure of the resulting residuals.
Fig 2: Spectral estimates on a log-log scale, with frequency units of cycles per year. The regression line is computed by regressing the log of multitaper estimator onto the log-frequencies.

We first explore the correlation structure of $\eta_t$ using estimates of the spectral density, $f(\lambda)$, of the empirical residuals. If the residuals have long-memory behavior then the logarithm of the spectrum will feature a negative slope with respect to log-frequency. More specifically, a stationary stochastic process $X_t$ is generally said to have long memory when its autocovariance function $\gamma(n) := \text{cov}(X_{t+n}, X_t)$ decays at the rate $n^{2H-2}$ for large time lag $n$, where $0.5 < H < 1$ is the long-memory parameter. This behavior is essentially equivalent to requiring that $f(\lambda)$ have a singular behavior $\lambda^{1-2H}$ for small frequencies $\lambda$ (see Beran, 1994). Since $1 - 2H < 0$ for long-memory models, the plot of $\log f(\lambda)$ against $\log \lambda$ for a long-memory model will be approximately a straight line with negative slope $1 - 2H$. While spectral methods are not generally accepted as a formal way to estimate $H$, save for very simple models, they do offer a useful diagnostic tool to evaluate the long-memory structure in the data (see Beran, 1994).

Based on the regression residuals from Eq.(3.2), we compute two widely used estimators of the spectral density: the periodogram and the adaptive multitaper estimator (see the Online Supplement ?? for a brief description for each estimator). Figure 2 shows both estimators on a log-log scale for the HAD and CRU dataset, respectively. In both cases, the multitaper spectral estimator features a clear negative slope on the log-log scale, indicating possible long-memory behaviors. Results for the periodogram are less striking than the multitaper estimate, but still show a negative slope in log-log.
To examine more formally the long memory behavior of the residuals, we employ the test developed by Robinson (1995) (see Section 3.4 for results of alternative tests). To introduce the idea of these methods briefly, consider a stationary process $X_t$ with spectral density $f(\lambda)$. The $f(\lambda)$ may satisfy the power law $f(\lambda) \sim G\lambda^{1-2H}$ as $\lambda \to 0$ for a positive value $G$ and some $H \in (0, 1)$. The so-called Hurst parameter $H$ measures the length of the correlation as illustrated by the negative slope of the spectrum in Figure 2. Typical examples that follow this power law include FARIMA and fGn. In fact, fGn is the discrete-time stationary Gaussian process that is the first-order difference process of the so-called fractional Brownian motion (fBm) process evaluated at integer times. The spectrum of the distributional derivative of the fBm process is proportional to $\lambda^{1-2H}$. The spectrum of fGn has the same behavior asymptotically for small $\lambda$. The fBm process is self-similar. While is not the case for the discrete time fGn, its parameter $H$ is sometimes called the self-similarity index nonetheless, because of the connection to fBm. Historically, the parameter $H$ first made its appearance when fBm was introduced by Kolmogorov (1940); the name Hurst arose after Mandelbrot proposed that fBm might be a good model to explain the power behavior of a statistic introduced by the hydrologist H.E. Hurst to study yearly levels of the Nile river: see Mandelbrot (1965); Mandelbrot and Van Ness (1968) and the account in Taqqu (2013). More information on fGn can be found in the Online Supplement ??.

The null hypothesis for the Robinson (1995) test is $H = 0.5$ (no memory), while the alternative hypothesis is $H > 0.5$ (long-memory). The test is based on the asymptotic normality of the semiparametric Gaussian estimate of $H$. Other tests are reviewed in Murphy and Izzeldin (2009), who recommend Robinson’s test due to its power properties and its good performance for relatively small samples when combined with bootstrap resampling. For long-memory models that tend towards only local self-similarity –which is not the case for fGn– the \( \tilde{H} \) used in Robinson (1995) may not be consistent for $H$, but the test remains valid.

We perform Robinson’s test on the regression residuals in (3.2), resulting in p-values of 0.0258 for HAD and 0.0002 for CRU. Thus both datasets show strong evidence in favor of rejecting the null hypothesis of $H = 0.5$.

3.3. Possible long-memory behavior in the temperature anomalies. In the specification of the process level of the hierarchical model, we follow Li, Ny-
chka and Ammann (2010) and model the latent temperature as linear in the external forcings. We apply the following transformations to the forcings, where $S$, $V$ and $C$ are, respectively, the time series of solar irradiance, volcanism and greenhouse gases:

- $\tilde{V}_t = \log(-V_t + 1)$. Exploratory data analysis indicated that this transformation increases the explanatory power of volcanism. From a physical standpoint, it dampens the effects of very large events, and thus provides a form of regularization given the larger uncertainties associated with the larger $V$ values (Li, Nychka and Ammann, 2010).
- $\tilde{C}_t = \log(C_t)$. Following the (Hegerl et al., 2007), we use a log-transformation to approximate the radiative forcing due to changes in the equivalent CO$_2$ concentration.

The resulting process-level model is,

$$T_t = \beta_0 + \beta_1 S_t + \beta_2 \tilde{V}_t + \beta_3 \tilde{C}_t + \sigma_T \epsilon_t,$$

where $\epsilon_t$ denotes a stationary stochastic process with zero mean and unit variance, and $\sigma_T$ is a constant variance parameter. Li, Nychka and Ammann (2010) employ an AR(2) for the error term, based on an examination of auto- and partial auto-correlation functions. However, in a similar situation, Beran (1994) shows that the residuals are appropriately modeled as FARIMA($0, d = 0.4$, 0), with Hurst parameter $H = d + 0.5 = 0.9$. Benth and Saltytė-Benth (2005) and Brody, Syroka and Zervos (2002) also provide examples of estimation of long-memory parameters over regression residuals on temperature series for specific locations in Norway and England, respectively, while Huybers and Curry (2006) provides statistical evidence of a power-law behavior in the spectrum of surface temperatures. Finally, Imbers et al. (2013) uses a long-memory fractional-differencing process that is very similar to fGn in terms of its asymptotic long-memory behavior, in order to test the presence of an anthropogenic impact on present-day temperatures.

We repeat the same diagnostic procedure and hypothesis testing as in Section 3.2 to assess the long memory behavior of $\epsilon_t$. We first fit model (3.3) using ordinary least-squares criterion, and find $R^2$ values of 73% for HAD and 66% for CRU, indicating the strong explanatory power of the forcings. Figure 3 plots spectral density estimates in log-log space, for both HAD and CRU, and shows that HAD, but not CRU, exhibits a negative slope. The p-value associated with Robinson’s test is $8.39 \times 10^{-7}$ for HAD and 0.058 for CRU, indicating a long-range correlation structure for the HAD data set but not for CRU.
3.4. Other tests. We briefly discuss results for several alternatives to Robinson’s test.

Beran’s test (See Beran, 1992) evaluates the goodness-of-fit of an fGn model to a stochastic process. Let $X_t$ be a stationary Gaussian process with spectral density $f(\lambda)$. If $f(\lambda, H)$ is the spectral density of an fGn process with Hurst parameter $H$, then the null hypothesis for Beran’s test is $H_0: f(\lambda) = f(\lambda, H)$ and the alternative is $H_a: f(\lambda) \neq f(\lambda, H)$. Both the Robinson and Beran tests base their test statistics on the Whittle estimator of $H$, which enjoys the desirable property of insensitivity to the scale of self-similarity (see the Online Supplement ?? for additional technical details).

We performed Beran’s test on all four residual datasets, under three null hypotheses: (a) fractional Gaussian noise, (b) AR(1), and (c) AR(2). The corresponding p-values are shown in Table 2, for residuals from the HAD and CRU datasets, for both the proxy [Eq. 3.2] and instrumental [Eq. 3.3] equations. Results indicate that Beran’s test cannot reject the null in any case.

Finally, we apply the test proposed by Davies and Harte (1987); see section ?? for technical details. The fGn is used as the underlying parametric model. The null and alternative hypotheses are identical to Robinson’s test. P-values in Table 3 show that the null can be rejected in three out of our four cases.

No single method employed here is a perfect indicator for the presence
Table 2

Results of Beran’s test applied to the the residuals from the HAD and CRU datasets, for both the proxy [Eq. 3.2] and instrumental [Eq. 3.3] equations, under three null hypotheses

| Model  | HAD-Proxy | CRU-Proxy | HAD-Temp | CRU-Temp |
|--------|-----------|-----------|----------|----------|
| fGn    | 0.78      | 0.91      | 0.56     | 0.73     |
| AR(1)  | 0.76      | 0.88      | 0.57     | 0.74     |
| AR(2)  | 0.83      | 0.93      | 0.57     | 0.74     |

Table 3

As in Table 2, but for the Davies and Harte test.

| Model  | HAD-Proxy | CRU-Proxy | HAD-Temp | CRU-Temp |
|--------|-----------|-----------|----------|----------|
| fGn    | 0.046     | 0.000     | 0.010    | 0.436    |

or absence of long-memory processes. Taken together, however, the spectral density estimates, and applications of the tests of Robinson (1995), Beran (1992), and Davies and Harte (1987) indicate to us that the possibility of long memory cannot be ignored in developing models for the residuals. As we will see, the Bayesian fits of our full model provide quantitative evidence in favor of long memory error process over noise processes with no memory, as indicated by the posterior distribution of the long memory parameters; see Section 4.1.

3.5. Hierarchical Bayesian model with long-memory errors. Given the statistical evidence for possible long memory correlation in the empirical residuals from Eqs (3.2) & (3.3), and the implication for fGn by Beran’s test, we may consider modeling the errors using either fGn or an AR process. Both models lead to a similar strategy for implementing a hierarchical Bayesian reconstruction. Since the case of fGn errors is more involved computationally, we choose to present this case in most detail. Comparisons between various modeling choices (long memory vs. with short memory vs with no memory; with or without forcings) are given in Section 4.2 (see e.g. Table 4). A summary of our hierarchical models in the data and process stages is as follows:

\[ \begin{align*}
RP_t & = \alpha_0 + \alpha_1 T_t + \sigma_P \eta_t, \\
T_t & = \beta_0 + \beta_1 S_t + \beta_2 \tilde{V}_t + \beta_3 \tilde{C}_t + \sigma_T \epsilon_t,
\end{align*} \tag{3.4} \]

where \( \eta_t \) and \( \epsilon_t \) are independent fGn processes with respective parameters \( H \in (0, 1) \) and \( K \in (0, 1) \) which control the long memory behavior. We assume these models hold throughout the entire prediction period (1000-1899) and calibration period (1900-1998). Independence between \( \epsilon_t \) and \( \eta_t \) is
a reasonable assumption as $\eta_t$ represents the stochastic aspect of the proxies that is not explained by the climate, while $\epsilon_t$ is the long-memory aspect of the climate not attributable to the forcings.

The modeling framework (Eq. 3.4) is based on the assumption that the relationship between the proxies and temperatures is invariant through time. While stationarity may be an idealized assumption, we note that our data selection procedure ensures that stationarity is at the very least not an unreasonable assumption, while the short calibration period precludes a more in-depth study of possible non-stationarity in the temperature–proxy relationship. Moreover, we note that the modeling framework could be made more realistic by specifying a (possibly independent) error structure for each individual proxy series. We do not pursue these specifics here, but rather focus on exploring the effects of long memory and forcings on the reconstruction.

Following Li, Nychka and Ammann (2010), we define the following prior distributions for the parameters $\alpha := (\alpha_0, \alpha_1)^T$, $\beta := (\beta_0, \beta_1, \beta_2, \beta_3)^T$, $\sigma_1^2$, $\sigma_2^2$, $H$ and $K$:

\begin{itemize}
  \item $\alpha \sim N((0, 1)^T, I_2)$; $\beta \sim N((0, 1, 1, 1)^T, I_4)$;
  \item $\sigma_1^2 \sim IG(2; 0.1)$, $\sigma_2^2 \sim IG(2; 0.1)$;
  \item $H \sim \text{Unif}(0, 1)$; $K \sim \text{Unif}(0, 1)$.
\end{itemize}

where $I_n$ is the identity matrix of dimension $n$.

Let $T_u = (T_{1000}, \ldots, T_{1899})$ denote the vector of unknown temperatures and $T_0 = (T_{1900}, \ldots, T_{1998})$ the vector of instrumental temperatures. Our goal is to infer $T_u$ based on $T_0$, $RP$, $S$, $\tilde{V}$ and $\tilde{C}$. The full conditional posterior distributions of $T_u$ and all unknown parameters save $H$ and $K$ can be derived explicitly, thus allowing for standard Gibbs sampling in the Markov chain Monte Carlo (MCMC) method. We resort to Metropolis-Hasting steps to sample $H$ and $K$. The derivation of full conditional distributions can be found in the Online Supplement. We implement the MCMC using a number of R packages: MCMCpack (Martin, Quinn and Park, 2011), mvtnorm (Genz and Bretz, 2009), ltsa (McLeod, Yu and Krougly, 2007) and msm (Jackson, 2011).

\section{Numerical results.}
For reconstructions using both the HAD and CRU instrumental records, we sample 5000 times from the posterior distribution and discard the first 1000 replicates to account for the burn-in period. The details of posterior samples are shown in the Online Supplement. Here we summarize the results and show a selection of representative plots and focus on reconstructions using the HAD data set.
4.1. Bayesian parameter estimates. Figure 4 shows trace plots and histograms of the \( H \) and \( K \) parameters that are responsible for long memory in the HAD reconstruction, for the data- and process-level residuals, respectively. Visually, the posterior draws quickly stabilize; see Section 4.3 for a formal assessment of convergence for these and other parameters. The histograms of \( H \) and \( K \) for the HAD reconstruction clearly indicate that both parameters are significantly greater than 0.5, suggesting that the data are consistent with a long-range correlation model. While this does not imply that a short-memory AR model would be misspecified in a frequentist sense, it is an indication that there may be advantages to choosing a long-memory model, particularly in a Bayesian methodology. One such advantage appears to be a more accurate evaluation of uncertainty using fGn than using AR(1) error model, as can be seen in Table 4 below. Figure 5 shows the posterior distribution of \( H \) and \( K \) for the CRU reconstruction. The distribution of \( H \) (memory structure of the proxy residuals) is similar to that arising from the HAD analysis, whereas the posterior distribution of \( K \) for the CRU analysis is centered on smaller values than for HAD, but still remains significantly greater than 0.5. The larger value of \( K \) for the HAD data set, which includes the oceans, is in-line with intuition, on account of the larger heat capacity of the oceans and resulting longer timescale of response to changes in the forcings.

Posteriors samples for the process-level regression coefficients (the \( \beta_i \)) show that the transformed volcanic and greenhouse gas forcing series are meaning-

\[ \begin{align*}
\text{(a) Traceplots} & \quad \text{(b) Histograms}
\end{align*} \]

Fig 4: Bayesian estimation of \( H \) and \( K \) based on HAD Dataset.
ful predictors of the temperature evolution for both HAD and CRU, while solar irradiance is less influential (Figures ?? and ??). In general, posterior distributions of the remaining parameters are similar when using the CRU data set; see the Online Supplement ?? for additional plots.

4.2. Temperature reconstruction results. The temperature reconstruction for the HAD dataset is shown in Figure 6, together with its 95% point-wise credible intervals and boxplots summarizing the shape of the corresponding posterior distributions. Boxplots are formed from the empirical posterior distribution of the reconstructed temperatures for each year, and are displayed every fifty years, at 1000, 1050, . . . , 1850. The reconstruction shows a slight downward trend during the period 1000-1899 (cf. Kaufman et al., 2009), and no maxima in the posterior distributions exceed the levels observed after approximately 1950. The reconstruction for the CRU dataset (see Figures ?? and ??), is qualitatively similar, but features higher variance due to the more variable CRU temperatures. Repeating the reconstructions with the single lacustrine record excluded from the reduced proxy leads to very similar reconstructions; see Fig. ??.

In order to evaluate our reconstruction, we use 1900-1998 as an in-sample validation period. Due to the limited number of available observations and the necessity of inferring the long-memory parameters, out-of-sample validation was not feasible. Figure 7 shows the posterior mean and 95% point-wise credible intervals for predictions using the HAD data, as well as the actual HAD observations. Overall, the predictions and observations are in good
PAELOCLIMATE RECONSTRUCTION

Fig 6: Temperature reconstruction (1000-1899) using the HAD dataset. Boxplots correspond to the posterior distributions at 50-year intervals. The box is formed among the upper and lower quartile of posterior temperature anomalies, and the Whiskers extend out to minimum and maximum values (represented with horizontal lines).

qualitative agreement with one another, with the predictions tracking the overall shape of the warming trend over the 20th century. Results are similar for the CRU data set (see Figure ??). Note that the reduced variability of the posterior mean as compared with the observations is akin to the predictions from a linear regression being less variable than the observations. A key advantage of a Bayesian analysis, such as that used here, is that, provided the process-level model assumptions are reasonable, the temporal variability of individual posterior draws will be similar to that of the actual climate, even while variability of the mean across them is attenuated (see Fig. 2 of Tingley and Huybers, 2010b, for further discussion).

We provide a quantitative assessment of our reconstructions using a number of statistical measures: squared bias (squared sample mean of differences between the posterior mean and the observed anomalies), variance (sample variance of the differences used in bias calculation), and root mean squared error (RMSE) that combines the squared bias and variance; empirical coverage probabilities (ECP) of the credible intervals at the 95% and 80% levels; Interval Scores (IS) at the 95% and 80% levels; and, since we obtain MCMC samples from the predictive distribution, the Continuous Ranked Probabi-
Fig 7: Temperature reconstruction (1900-1998) using the HAD dataset.

bility Score (CRPS). The ECP measures the accuracy of the uncertainty quantification, while the IS and CRPS provide more nuanced assessments of the posterior predictive distributions, rewarding both the calibration and the sharpness simultaneously; further details of scoring rules are available in Gneiting and Raftery (2007); Gneiting, Balabdaoui and Raftery (2007); Gschlößl and Czado (2007), and Online Supplement ?? . For convenience, we report the negative IS and CRPS so the smaller means the higher quality of the predictions.

The novel aspects of our reconstructions are the incorporation of long-memory errors and external forcings, and we assess their benefits by considering six separate reconstruction scenarios:

SA: Possible long memory ($H$ and $K$ not fixed), with external forcings.
SB: AR(1) errors in formula (3.4), with external forcings.
SC: No memory ($H = K = \frac{1}{2}$), with external forcings.
SD: Possible long memory ($H$ and $K$ not fixed), no external forcings ($\beta_i = 0, i = 1, 2, 3$).
SE: AR(1) errors in formula (3.4), no external forcings ($\beta_i = 0, i = 1, 2, 3$).
SF: No memory ($H = K = \frac{1}{2}$), no external forcings ($\beta_i = 0, i = 1, 2, 3$).

We refer to scenarios C and F as having no memory, as they are based on Gaussian white noise errors that are independent and thus have no memory. Scenario A corresponds to the results discussed above and is our primary
focus, as it includes both long memory and forcings. We include Scenario B as the error structure features short memory, but more persistence than Scenarios C and F, and is therefore a useful intermediate for the purposes of comparison.

### Table 4
Validation measures of six scenarios for HAD and CRU datasets.

| Scenarios | Sq. Bias | Variance | RMSE | ECPβ95 | ECPβ80 | ISβ95 | ISβ80 | CRPS |
|-----------|----------|----------|------|--------|--------|-------|-------|------|
| **HAD**   |          |          |      |        |        |       |       |      |
| A         | .016     | .012     | .168 | 92.9   | 74.7   | 0.06  | 0.18  | 0.21 |
| B         | .015     | .011     | .162 | 90.9   | 74.7   | 0.06  | 0.18  | 0.21 |
| C         | .014     | .010     | .154 | 90.9   | 69.7   | 0.06  | 0.17  | 0.19 |
| D         | .055     | .072     | .356 | 99.0   | 84.8   | 0.11  | 0.32  | 0.24 |
| E         | .081     | .071     | .390 | 94.9   | 75.8   | 0.12  | 0.39  | 0.26 |
| F         | .113     | .059     | .415 | 82.8   | 59.6   | 0.17  | 0.51  | 0.36 |
| **CRU**   |          |          |      |        |        |       |       |      |
| A         | .032     | .025     | .238 | 91.9   | 73.7   | 0.08  | 0.25  | 0.24 |
| B         | .035     | .023     | .242 | 89.9   | 69.7   | 0.09  | 0.27  | 0.25 |
| C         | .031     | .024     | .235 | 91.9   | 73.7   | 0.09  | 0.25  | 0.24 |
| D         | .089     | .097     | .432 | 97.0   | 78.8   | 0.13  | 0.42  | 0.27 |
| E         | .120     | .095     | .464 | 90.9   | 75.8   | 0.15  | 0.48  | 0.30 |
| F         | .148     | .080     | .477 | 84.8   | 62.6   | 0.21  | 0.57  | 0.34 |

RMSE: Root Mean Square error, ECPβ: Empirical Coverage Probability at β% confidence level, ISβ: Interval Score at β% confidence level, CPRS: Continuous Ranked Probability Score.

* HAD and CRU refer to the two instrumental data sets, with HAD including the oceans.

Table 4 summarizes the quantitative assessments of the reconstructions for both the HAD and CRU datasets. The benefit of the external forcings are readily apparent (cf Li, Nychka and Ammann, 2010), as their inclusion substantially reduces the squared bias, variance, and consequently the RMSE (compare Scenario A to D, Scenario B to E and Scenario C to F). Moreover, the widths of the 95% credible intervals are likewise narrower when the external forcings are included (see Figure 8, below, and Figure ??). The RMSE is larger in Scenario A than C, with a larger fractional increase in variance (0.002/0.01=20%) than in bias (0.002/0.014=14.29%), and these differences are quantitatively similar under repeated sampling. Hence the inclusion of long-memory parameters in the Bayesian model has a larger impact on the variance than the squared bias, and results are similar for the CRU reconstruction. We note that the change in variance is close to what one might expect for theoretical reasons. Indeed, the increase of 15-20% is consistent with the increase in asymptotic variance in the central limit theorem for long-memory stationary sequences compared to the classical central limit theorem (see Theorem 7.2.4 in Nourdin and Peccati, 2012). The squared bias, variance, and RMSE are generally larger for the CRU reconstructions, which is based on a noisier data set, than for the HAD reconstructions (compare Figures 7 and ??). Comparing the ECPs between Scenarios A and C shows that the ac-
tual coverage rates are closer to their nominal rates when long memory is included. Thus the slight increase in variance seen in the long-memory scenarios has the desirable effect of resulting in more precise uncertainty quantification. Note in particular that the ECPs under the no-memory Scenario C are lower than their nominal values, indicating an underestimate of uncertainty in the absence of the long memory models. The increase in variability from Scenario C to A can also be explained directly in the Bayesian analysis, due to the relationship between the posterior mean of $\sigma_T$ and the variable $K$: we observe that if the temperature has long memory then the posterior mean of its variance is greater than is obtained when $K = 0.5$ (derivation included in Online Supplement ??).

For both the HAD and CRU data sets, inclusion of the forcings substantially decreases the IS and CRPS, pointing to the importance of forcings in constructing past temperatures (cf Li, Nychka and Ammann, 2010). The changes in the sharpness measures, $IS_{95}$ and $IS_{80}$, and the CRPS, are much smaller when comparing Scenario A, B, and C, than when comparing any of those scenarios with forcings, to the Scenario D, E and F which have no forcings, suggesting that the quality of the posterior distribution is less affected by the memory structure of the error model choice than by the forcings. Note, however, that when the forcings are not included, the long-memory error model reduces both the IS and CRPS scores.

Given the model assessments in Table 4, and the posterior estimates of model parameters, we argue that Scenario A is the most appropriate model for the data. The substantive deviations between the ECPs and their nominal coverage rates for Scenarios D, E and F, for both the HAD and CRU reconstructions, suggest that the models that forego the forcings are misspecified. In addition, the posterior distributions of the coefficients for both the volcanic and greenhouse gas forcing series are significant, providing further evidence that the forcings should be included.

The differences in validation metrics between Scenarios A, B and C, all of which include the external forcings, are minor as compared with the difference between any of those scenarios and those without the forcings. Indeed, tests for selecting between long and short memory models for climate time series are often inconclusive (e.g. Percival, Overland, and Mofjeld, 2001). For both the HAD and CRU reconstructions, the ECPs are generally closest to their nominal rates for Scenario A, while the proper scoring rules are similar between the three scenarios that include the forcings. Since the error structure mainly affects the uncertainty quantification by increasing the variance, we view Scenario A as a modeling approach that is conservative from the perspective of reporting uncertainty. Finally, the posterior estimate of the
long memory parameters, for both the HAD and CRU data sets (see Figures 4 and 5) clearly indicate the appropriateness of modeling the residuals with long memory processes.

4.3. **MCMC Diagnostics.** To establish convergence of the MCMC samples, we examine trace plots (Figures ??–??), and calculate the potential scale reduction factor (PSRF; Gelman and Rubin, 1992) and its multivariate version (Brooks and Gelman, 1998); see Brooks and Roberts (1997) and Cowles and Carlin (1996) for further details. If the PSRF is close to unity for all parameters, then the Markov chain simulation is close to its stationary distribution, while a large PSRF indicates that the chain has not converged (Gelman and Rubin, 1992). Brooks and Gelman (1998) provide a generalization that allows for the computation of a single PSRF for all model parameters.

For both the HAD and CRU datasets, we run five MCMC simulations, each of length 5000, and discard the first 1000 samples to allow the chain to burn in. We compute PSRFs for the scalar parameters of the model \((\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2, \beta_3, \sigma_1^2, \sigma_2^2, H, K)\) and the multivariate PSRF, along with their upper 95% confidence bounds, using the \texttt{coda} R-package (Plummer et al., 2006). Results in Table 5 show that all the individual PSRFs are relatively close to unity, indicating their successful convergence to the stationary distribution. The multivariate PSRF likewise indicates convergence.

|        | \(\alpha_0\) | \(\alpha_1\) | \(\beta_0\) | \(\beta_1\) | \(\beta_2\) | \(\beta_3\) | \(\sigma_1^2\) | \(\sigma_2^2\) | \(H\) | \(K\) | Mul. |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------|------|------|
| HAD PSRF | 1.01        | 1.01        | 1.01        | 1.01        | 1.00        | 1.01        | 1.00        | 1.01        | 1.01 | 1.01 | 1.02 |
| UB     | 1.02        | 1.04        | 1.01        | 1.03        | 1.02        | 1.00        | 1.02        | 1.01        | 1.01 | 1.03 | –    |
| CRU PSRF | 1.00        | 1.01        | 1.03        | 1.04        | 1.01        | 1.00        | 1.01        | 1.00        | 1.00 | 1.00 | 1.05 |
| UB     | 1.00        | 1.03        | 1.09        | 1.11        | 1.01        | 1.02        | 1.01        | 1.01        | 1.00 | 1.01 | –    |

5. Comparison with other works.

5.1. **Comparison with previous reconstructions.** We compare our reconstructions to those reported in Mann et al. (2008a), as both use similar proxy and temperature data sets. Mann et al. (2008a) do not include long memory error processes or the external forcings, and present reconstructions, along with uncertainty bands, based on two regression approaches: composite plus scale (CPS) and errors in variables (EIV). The CPS approach computes a weighted average of the proxy data, and then calibrates this weighted average
by matching its mean and variance to those of the instrumental temperature data during their overlap period. The EIV regression approach allows for errors in both the dependent and independent variables, and we refer to Mann et al. (2008a,b) for details. The EIV and CPS reconstructions, and their associated uncertainty estimates, are available on-line as decadalely smoothed time series, as Mann et al. (2008a) focusses on low-frequency climate variability. In contrast, the reconstructions we present here are available at annual temporal resolution, with no smoothing. In comparisons, we show the posterior mean and uncertainty of our reconstructions at annual resolution, and additionally include the posterior mean that results from first smoothing each posterior draw with a Butterworth filter with cutoff frequency equal to 0.1 cycles/year.

Figure 8 compares our reconstructions, using the HAD data and under the six modeling scenarios discussed above, to those from Mann et al. (2008a). In all cases, and especially when including the forcings, our reconstructions are generally cooler than both the EIV and CPS reconstructions from Mann et al. (2008a), particularly during the 1000–1400 interval, and feature a smaller amplitude of pre-instrumental temperature variability. We are not the first to report a lower variability than Mann et al. (2008a) – for example, PAGES 2k Consortium (2013) report a change in 30 year average temperatures between 1000AD and the 1800s of about 0.3°C, compared with about 0.5°C for Mann et al. (2008a); see Fig. 4 of PAGES 2k Consortium (2013). Qualitatively, our results are indicative of even lower variability on multi-decadal intervals (e.g., the box plot centers plotted in Fig. 6).

The model settings of Mann et al. (2008a) are most similar to our Scenario F, which includes neither the forcings nor the long memory processes. Indeed, the EIV predictions from Mann et al. (2008a) are visually most similar to smoothed Scenario F results, and 88.4% of the EIV predictions from Mann et al. (2008a) fall within the 95% point-wise credible intervals for the smoothed Scenario F results. Results are similar when using the CRU observations (Figure ??).

To facilitate numerical comparisons with the Mann et al. (2008a) reconstruction, we re-calculate validation metrics for Scenario F after first smoothing each posterior draw; results are shown in Table 6 for the 20th century validation interval. The main difference between our smoothed Scenario F results and the Mann et al. (2008a) results is in terms of squared bias, with

\[ \text{http://www.ncdc.noaa.gov/paleo/pubs/mann2008/mann2008.html} \]

\[ \text{Our calculations are based on the Matlab code associated with Mann et al. (2008a), posted on one at http://www.ncdc.noaa.gov/paleo/pubs/mann2008/mann2008.html. We smooth using the filtfilt command in the R package “signal”} \]
the Mann et al. (2008a) reconstruction featuring biases that are about an order of magnitude smaller, and variances that are about 1.5–2 times larger. The net result is that the Mann et al. (2008a) reconstructions feature smaller RMSE than our smoothed Scenario F, on par with results from our annually resolved Scenario A. As measured by the ECP, the uncertainties for the Mann et al. (2008a) reconstructions are too wide, in the sense that the
empirical coverage rate is greater than the nominal rate. The uncertainties for our smoothed Scenario F is smaller than that in Mann et al. (2008a), but due to the relatively large bias, the ECPs appear to be too low compared to their nominal value. The Interval Scores for the smoothed Mann et al. (2008a) reconstructions are much better than those for our smoothed Scenario F, and like the RMSE, are similar to those for our annually resolved Scenarios A, B and C, which carry small squared bias by including the forcings (see Table 4).

We caution against drawing substantive conclusions from the comparison of the validation and scoring metrics between the Mann et al. (2008a) results and the smoothed Scenario F, as numerous lines of evidence indicate that F is the least appropriate of the six scenarios explored here: validation metrics and scores (Table 4) are generally the worst for Scenario F; the inclusion of the forcings is motivated by the scientific understanding of their connection with temperatures; and the inclusion of the long memory processes is driven by the structure of the data. As discussed above, we view Scenario F as a mis-specified model, and the high squared bias and associated inadequacies of the ECPs are therefore to be expected. Perhaps the most telling conclusion to be drawn from the numerical comparisons is that our annually resolved Scenario A, which includes the forcings and long memory processes, is comparable in terms of RMSE and Interval Scores to the decadally resolved Mann et al. (2008a) results.

Finally, we note that the proxy selection and modeling treatments do differ between our Scenario F and the reconstructions in Mann et al. (2008a) so that the comparison remains imperfect. In particular, we note that the Mann et al. (2008a) reconstructions includes proxies with decadal resolution,

| Scenarios | Sq. Bias | Variance | RMSE | ECP95 | ECP80 | IS95 | IS80 |
|-----------|----------|----------|------|-------|-------|------|------|
| HAD       | F (smoothed) | .100     | .012 | .335  | 41.4  | 33.3 | .46  | .71  |
|           | CPS      | .009     | .024 | .183  | 100.0 | 96.9 | .06  | .16  |
|           | EIV      | .003     | .022 | .157  | 99.0  | 99.0 | .06  | .23  |
| CRU       | F (smoothed) | .121     | .016 | .371  | 48.5  | 36.4 | .45  | .73  |
|           | CPS      | .017     | .025 | .207  | 99.0  | 99.0 | .07  | .25  |
|           | EIV      | .006     | .021 | .163  | 98.0  | 98.0 | .07  | .17  |

*The statistics for EIV and CPS reconstructions are calculated using the estimated standard deviations associated with Mann et al. (2008a). They are posted as “2-sigma uncertainties” (S), hence the formula for their 95% confidence bands is $M_t \pm 1.96\ S$, where $M_t$ is their predicted temperature mean.
whereas here we focus on proxies with annual resolution. Indeed, the CPS reconstruction is performed after smoothing all proxies to a common decadal resolution, while the EIV reconstruction is based on a “hybrid” frequency approach that involves separate calibrations to infer climate on interdecadal (periods longer than 20 years) and interannual (periods shorter than 20 years) timescales (Mann et al., 2008b, 2007). Due to the differing methods and the focus on lower frequency variability in Mann et al. (2008a), the differing validation metrics between our Scenario F and those for the Mann et al. (2008a) reconstructions are not surprising.

5.2. Sensitivity Measure. The Fourth Assessment Report of the IPCC (see p. 723 in Hegerl et al., 2007) refers to the “transient climate response” (TCR) as the “global mean temperature change that is realized at the time of CO$_2$ doubling . . . TCR is therefore indicative of the temperature trend associated with external forcing, and can be constrained by an observable quantity, the observed warming trend that is attributable to greenhouse gas (GHG) forcing”. In our model, the transient response to a doubling of GHG is embedded in the parameter $\beta_3$, and the resulting estimates of TCR are based on observed temperature increases since 1900, and proxy and forcing information over the past millennium. We believe that our Bayesian approach to computing the transient response to GHG forcing from both instrumental and proxy observations, without recourse to global climate models, is new to the field.

Taking into account the transformations applied to the CO$_2$ series, we define TCR in terms of $\beta_3$ as:

$$\text{TCR} := \beta_3 \log 2 / \sigma(\log C),$$

where $\sigma(\log C)$ is the standard deviation of the logarithm of the GHG series $C$, and is computed over the entire period 1000-1998. An important advantage of Bayesian estimation is the possibility of obtaining a sample estimate of the marginal posterior distribution of $\beta_3$ given the data. From this information we can compute a non-parametric estimator of the probability density function for TCR, as shown in Figure 9, which accounts for the uncertainties in all the other parameters in the model.

Our estimate of TCR using the HAD dataset shows a median around $2.56^\circ$C, with an approximate 95% credible interval of $[2.20, 2.95]^\circ$C. The median is in good agreement with the land-and-marine TCRs reported in Figure 9.21 in Hegerl et al. (2007), while our spread is significantly less than the range of $1.5 – 2.8^\circ$C (all uncertainty intervals quoted from the literature are 95% unless otherwise stated) from Hegerl et al. (2007). We note that the
IPCC estimates are based on global climate models, whereas our methodology is exclusively data-based. A possible cause for the narrower uncertainty is the more extensive use of data, in terms of both variety (instrumental temperatures, CO$_2$, and proxies) and duration (observations over the last millennium).

Several recent studies have arrived at TCR estimates by combining information from models and the instrumental temperature record. Gillett et al. (2012) produce a TCR estimate of 1.3 – 1.8°C using the global HAD data set and a single global climate model, but note that this TCR estimate may be unrealistically narrow as it results from a single climate model. A more recent study (Gillett et al., 2013) that combines information from an ensemble of models and the instrumental record results in wider range of TCR estimates, 0.9 – 2.3°C, featuring greater overlap with our results. Otto et al. (2013) use global, decadal averages of the HAD data set over the 1970–2009 to arrive at a data-based TCR estimate in the range of 0.7 – 2.5°C, but caution against strong conclusions based on a such a short time interval. The recently released IPCC Fifth Assessment Report (AchutaRao et al., 2013) summarizes these and other studies to arrive at an expert assessment that TCR is likely in the range 1 – 2.5°C, and is extremely unlikely to be greater than 3.0°C. Both the specific model-data fusion studies discussed here, and the synthesis provided by the IPCC Fifth Assessment report, are generally on the low end of the 95% posterior credible interval we derive for TCR,
likely a result of the longer time span and different data types that enter into our estimate.

The land-only TCR distribution from the CRU data set is slightly wider than that for the land-and-marine HAD data set, and feature a median that is about 0.4°C larger. That the land features a bigger response is reasonable, as it has lower heat capacity than the ocean, and thus should respond faster and feature a higher TCR. The slight increase in variability is inherited from the difference in the posterior variability of \( \hat{\beta}_3 \), as seen by comparing Figures ?? and ?? in the Online Supplement ??.

6. Conclusions and Discussion. We use a comprehensive multiproxy data set to produce reconstructions of the NH temperature anomaly time series back to 1000 AD. Our reconstructions exploit the information available in climate forcings, and explores the role of possible short and long memory processes in the residuals. Hierarchical modeling provides a natural framework for integrating the different information sources (proxy and instrumental temperatures observations, and time series of solar, greenhouse gas, and volcanic forcings), while Bayesian inference allows for estimation of all unknown quantities, including past temperatures, and facilitates uncertainty propagation.

The possibility of long-memory is suggested by exploratory data analysis using traditional statistical assessment techniques, and is reinforced by the Bayesian estimation result. Nevertheless, short-memory models such as AR(1) can still provide comparable results as those obtained with long-memory models using the current data set. However, allowing for the long-range correlation in the model may be considered a more conservative approach to quantifying uncertainty. The inclusion of the external forcings is motivated from physical principles and the conclusions of Li, Nychka and Ammann (2010), and additionally allows for estimation of the transient climate response. While our TCR distribution is on the high side of recently published ranges, we note that our estimate is based on both the instrumental and paleoclimate records, and does not rely on global climate models.

By considering six modeling scenarios, where the external forcings are each either included or excluded and the residual processes are modeled using either long or short memory processes, we evaluate the effects that these modeling choices have on the reconstructions. Results show that allowing for long memory residuals improves the accuracy of uncertainty quantification, while using external forcings results in substantial reductions in both reconstruction bias and variability. The scenario with neither forcings nor memory is somewhat similar to the benchmark reconstruction of Mann et al. (2008a),
though we note that there remain differences in both method and data usage. Our reconstructions generally indicates cooler temperatures than those of Mann et al. (2008a), particularly before the year 1400.

The basic framework presented in this paper can be extended in several directions, and we anticipate that doing so will produce further insights into the climate of the late Holocene. An obvious extension is to incorporate a spatial element, by combining the model used here with the space-time model in Tingley and Huybers (2010a). Doing so would require generalizing the reduced-proxy framework used here, and instead specifying a separate long memory error model for each proxy time series, or perhaps a common model for each proxy type (c.f., Tingley and Huybers, 2010a). Such an implementation would pose technical challenges, as the estimation of the long memory parameters is the most numerically demanding component of the analysis. Prior scientific understanding of the mechanisms by which the proxies record variations in the climate may be helpful in selecting appropriate temporal correlation models for the residuals, and can potentially be used to simplify calculations. Such a computationally demanding generalization may be a more scientifically defensible use of the proxies, and may allow for further insights into the proxy–climate relationship. Lastly, major volcanic events occur sparsely and unpredictably in time, and these features are not shared by the other forcings. It may therefore be beneficial to consider a different type of time series model for the volcanic forcing time series, and, given the fixed-time-step context of the analysis, the classical binomial approximation to the Poisson process could provide an appropriate building block.

References.
AchutaRao, K. M., Allen, M., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I., Overland, J., Perlwitz, J., Sebbari, R. and Zhang, X. (2013). Detection and Attribution of Climate Change: from Global to Regional. 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 (J. Bartholy, R. Vautard and T. Yasunari, eds.) To be published.
Alexander, L., Allen, S., Bindoff, N., Breon, F.-M., Church, J., Cubasch, U., Emori, S., Forster, P., Friedlingstein, P., Gillett, N., Gregory, J., Hartmann, D., Jansen, E., Kirtman, B., Knutti, R., Kumar Kanikcharla, K., Lemke, P., Marotzke, J., Masson-Delmotte, V., Meehl, G., Mokhov, I., Piao, S., Ramaswamy, V., Randall, D., Rhein, M., Rojas, M., Sabine, C., Shindell, D., Talley, L., Vaughan, D. and Xie, S.-P. (2013). Technical Summary. 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 (S. Joussaume, J. Penner and F. Tangang, eds.) To be published.
Ammann, C. M., Joos, F., Schmel, D. S., Otto-Bliesner, B. L. and Tomas, R. A. (2007). Solar influence on climate during the past millennium: results from transient
simulations with the NCAR climate system model. *Proceedings of the National Academy of Sciences* 104 3713–3718.

Benth, F. E. and Šaltytė-Benth, J. (2005). Stochastic modelling of temperature variations with a view towards weather derivatives. *Applied Mathematical Finance* 12 53–85.

Beran, J. (1992). A goodness-of-fit test for time series with long range dependence. *Journal of the Royal Statistical Society. Series B (Methodological)* 54 749–760.

Beran, J. (1994). *Statistics for long-memory processes*. Monographs on Statistics and Applied Probability, 61. Chapman & Hall.

Berliner, L. M., Wikle, C. K. and Cressie, N. (2000). Long-lead prediction of pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* 13 3953–3968.

Briffa, K. R., Jones, P. D., Schweingruber, F. H. and Osborn, T. J. (1998). Influence of volcanic eruptions on Northern Hemisphere summer temperature over the past 600 years. *Nature* 393 450–455.

Brody, D. C., Syroka, J. and Zervos, M. (2002). Dynamical pricing of weather derivatives. *Quantitative Finance* 2 189–198.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B. and Jones, P. D. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research* 111 D12106+.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7 434–455.

Christiansen, J. H., Hewittson, B., Busuioc, A., Chen, A., Gao, X., Held, I., Jones, R., Kolli, R. K., Kwon, W. T., Laprise, R., Magaña Rueda, V., Mearns, L., Menéndez, C. G., Räisänen, J., Rinke, A., Sarr, A. and Whetton, P. (2007). Regional climate projections. 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* (S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller, eds.) Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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

Christiansen, B., Schmith, T. and Thejll, P. (2009). A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness. *Journal of Climate* 22 951–976.

Chronopoulou, A. and Viens, F. (2009). Hurst Index estimation for self-similar processes with long-memory. In *Recent development in stochastic dynamics and stochastic analysis* (J. Duan, S. Luo and C. Wang, eds.). Interdisciplinary Mathematical Sciences 8 6 91–117. World Scientific Publishing Co.

Chronopoulou, A., Viens, F. G. and Tudor, C. A. (2009). Variations and Hurst index estimation for a Rosenblatt process using longer filters. *Electronic Journal of Statistics* 3 1393–1435.

Chronopoulou, A. and Viens, F. (2012). Estimation and pricing under long-memory stochastic volatility. *Annals of Finance* 8 379–403.

Cook, E. R., Briffa, K. R. and Jones, P. D. (1994). Spatial regression methods in dendroclimatology: A review and comparison of two techniques. *International Journal of Climatology* 14 379–402.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 91 883–904.
Crowley, T. J., Criste, T. A. and Smith, N. R. (1993). Reassessment of Crete (Greenland) ice core acidity/volcanism link to climate change. Geophysical Research Letters 20 209–212.

Crowley, T. and Kim, K. (1993). Towards development of a strategy for determining the origin of decadal-centennial scale climate variability. Quaternary Science Reviews 12 375–385.

Crowley, T. J. and Kim, K. (1996). Comparison of proxy records of climate change and solar forcing. Geophysical Research Letters 23 359–362.

Davies, R. B. and Harte, D. S. (1987). Tests for Hurst effect. Biometrika 74 95–101.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 1–38.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science 7 457–472.

Genz, A. and Bretz, F. (2009). Computation of multivariate normal and t probabilities. Lecture Notes in Statistics, 195. Springer.

Gillett, N. P., Arora, V. K., Flato, G. M., Scinocca, J. F. and von Salzen, K. (2012). Improved constraints on 21st-century warming derived using 160 years of temperature observations. Geophysical Research Letters 39 L01704.

Gillett, N. P., Arora, V. K., Matthews, D. and Allen, M. R. (2013). Constraining the ratio of global warming to cumulative CO2 emissions using CMIP5 simulations. Journal of Climate 26 6844–6858.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 359–378.

Gneiting, T. and Schlather, M. (2004). Stochastic models that separate fractal dimension and the Hurst effect. SIAM Review 46 269–282.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. Scandinavian Actuarial Journal 2007 202–225.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B. and Mitchell, F. J. G. (2006). Bayesian palaeoclimate reconstruction. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169 395–438.

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, Y., Marengo Orsini, J. A., Nicholls, N., Penner, J. E. and Stott, P. A. (2007). Understanding and attributing climate change. 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 (S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Avert, M. Tignor and H. L. Miller, eds.) Cambridge University Press.

Huybers, P. and Curry, W. (2006). Links between annual, Milankovitch and continuum temperature variability. Nature 441 329–332.

Imbers, J., Lopez, A., Huntingford, C. and Allen, M. (2013). Sensitivity of climate change detection and attribution to the characterization of internal variability. Journal of Climate. To be published.

Jackson, C. (2011). Multi-state models for panel data: The msm package for R. Journal of Statistical Software 38 1–28.

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., Luterbacher, J., Wahl, E. R., Zwiers, F. W., Mann, M. E.,
PALEOCLIMATE RECONSTRUCTION

Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Janssen, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wolff, E. and Xoplaki, E. (2009). High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. The Holocene 19 3–49.

Juckes, M. N., Allen, M. R., Briffa, K. R., Esper, J., Hegerl, G. C., Moberg, A., Osborn, T. J., Weber, S. L. and Zorita, E. (2006). Millennial temperature reconstruction intercomparison and evaluation. Climate of the Past Discussions 2 1001–1049.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. Journal of Basic Engineering 83 95–108.

Kaufman, D. S., Schneider, D. P., McKay, N. P., Ammann, C. M., Bradley, R. S., Briffa, K. R., Miller, G. H., Otto-Bliesner, B. L., Overpeck, J. T., Vinther, B. M. and Others (2009). Recent warming reverses long-term arctic cooling. Science 325 1236–1239.

Kolmogorov, A. N. (1940). Wienersche spiralen und einige andere interessante kurven im Hilbertschen raum. In CR (Dokl.) Acad. Sci. URSS 26 115–118.

Landrum, L., Otto-Bliesner, B. L., Wahl, E. R., Conley, A., Lawrence, P. J., Rosenbloom, N. and Teng, H. (2013). Last millennium climate and its variability in CCSM4. Journal of Climate 26 1085–1111.

Lean, J., Beer, J. and Bradley, R. (1995). Reconstruction of solar irradiance since 1610: implications for climate change. Geophysical Research Letters 22 3195–3198.

Lee, T. C. K., Zwiers, F. W. and Tsao, M. (2008). Evaluation of proxy-based millennial reconstruction methods. Climate Dynamics 31 263–281.

Li, B., Nychka, D. W. and Ammann, C. M. (2010). The value of multiproxy reconstruction of past climate. Journal of the American Statistical Association 105 883–895.

Loso, M. G. (2009). Summer temperatures during the Medieval Warm Period and Little Ice Age inferred from varved proglacial lake sediments in southern Alaska. Journal of Paleolimnology 41 117–128.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M. and Wanner, H. (2004). European seasonal and annual temperature variability, trends, and extremes since 1500. Science 303 1499–1503.

Mandelbrot, B. B. (1965). Une classe de processus stochastiques homothétiques à soi: application à la loi climatologique de HE Hurst. Comptes Rendus Acad. Sci. Paris 240 3274–3277.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. SIAM Review 10 422–437.

Mann, M. E., Bradley, R. S. and Hughes, M. K. (1998). Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392 779–787.

Mann, M. E., Bradley, R. S. and Hughes, M. K. (1999). Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. Geophysical Research Letters 26 759–762.

Mann, M. E., Rutherford, S., Wahl, E. and Ammann, C. (2005). Testing the fidelity of methods used in proxy-based reconstructions of past climate. Journal of Climate 18 4097–4107.

Mann, M. E., Rutherford, S., Wahl, E. and Ammann, C. (2007). Robustness of proxy-based climate field reconstruction methods. Journal of Geophysical Research 112 D12109+.

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

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S. and Ni, F. (2008b). Supporting information for “Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia”. *Proceedings of the National Academy of Sciences* 105 13252–13257. [http://www.pnas.org/content/suppl/2008/09/02/0805721105.DCSupplemental0805721105SI.pdf](http://www.pnas.org/content/suppl/2008/09/02/0805721105.DCSupplemental0805721105SI.pdf).

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

Martin, A. D., Quinn, K. M. and Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software* 42 1–21.

McLeod, A. I., Yu, H. and Krougly, Z. L. (2007). Algorithms for linear time series analysis: with R package. *Journal of Statistical Software* 23 1–26.

McShane, B. B. and Wyner, A. J. (2011). A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable? *The Annals of Applied Statistics* 5 5–44.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., Raper, S. C. B., Watterson, I. G., Weaver, A. J. and Zhao, Z. C. (2007). Global climate projections. 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* (S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller, eds.) Cambridge University Press.

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

Murphy, A. and Izzeldin, M. (2009). Bootstrapping long memory tests: some Monte Carlo results. *Computational Statistics & Data Analysis* 53 2325–2334.

Nourdin, I. and Peccati, G. (2012). *Normal approximations with Malliavin calculus: from Stein’s method to universality* 192. Cambridge University Press.

NRC (2006). *Surface temperature reconstructions for the last 2000 years*. The National Academies Press, Washington, D.C.

Otto, A., Otto, F. E. L., Boucher, O., Church, J., Hegerl, G., Forster, P. M., Gillett, N. P., Gregory, J., Johnson, G. C., Knutti, R., Lewis, N., Lohmann, U., Marotzke, J., Myhre, G., Shindell, D., Stevens, B. and Allen, M. R. (2013). Energy budget constraints on climate response. *Nature Geoscience* 6 415–416.

Overpeck, J., Hughen, K., Hardy, D., Bradley, R., Case, R., Douglas, M., Finney, B., Gajewski, K., Jacoby, G. C. and Jennings, A. (1997). Arctic environmental change of the last four centuries. *Science* 278 1251.

PAGES 2k Consortium (2013). Continental-scale temperature variability during the past two millennia. *Nature Geoscience* 6 339–346.

Percival, D. B., Overland, J. E. and Mofjeld, H. O. (2001). Interpretation of North Pacific variability as a short-and long-memory process. *Journal of Climate* 14 7–11.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6 7–11.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics* 23 1630–1661.
Rutherford, S., Mann, M. E., Delworth, T. L. and Stouffer, R. J. (2003). Climate field reconstruction under stationary and nonstationary forcing. *Journal of Climate* 16 462–479.

Rutherford, S., Mann, M. E., Osborn, T. J., Briffa, K. R., Jones, P. D., Bradley, R. S. and Hughes, M. K. (2005). Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain. *Journal of Climate* 18 2308–2329.

Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* 14 853–871.

Smerdon, J. E., Kaplan, A., Chang, D. and Evans, M. N. (2010). A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium. *Journal of Climate* 23 4856–4880.

Smith, R. L. (2010). Understanding Sensitivities in Paleoclimatic Reconstructions. *Preprint.*

Steig, E. J., Schneider, D. P., Rutherford, S. D., Mann, M. E., Comiso, J. C. and Shindell, D. T. (2009). Warming of the antarctic ice-sheet surface since the 1957 international geophysical year. *Nature* 457 459–462.

Taqqu, M. S. (2013). Benoît Mandelbrot and fractional Brownian motion. *Statistical Science* 28 131–134.

Tiljander, M., Saarnisto, M., Ojala, A. E. K. and Saarinen, T. (2003). A 3000-year palaeoenvironmental record from annually laminated sediment of lake Korttajärvi, central Finland. *Boreas* 32 566–577.

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

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

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

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E. and Rajaratnam, B. (2012). Piecing together the past: statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews* 35 1–22.

Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A., Parker, D., Rahimzadeh, F., Renwick, J. A., Rusticucci, M., Soden, B. and Zhai, P. (2007). Observations: surface and atmospheric climate change. 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* (S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller, eds.).

Tudor, C. A. and Viens, F. G. (2007). Statistical aspects of the fractional stochastic calculus. *The Annals of Statistics* 35 1183–1212.

Wahl, E. R. and Smerdon, J. E. (2012). Comparison of paleoclimate field and index reconstructions derived from climate proxies and noise-only predictors. *Geophysical Research Letters* 39 L06703.

Werner, J. P., Luterbacher, J. and Smerdon, J. E. (2013). A pseudoproxy evaluation of Bayesian hierarchical modeling and canonical correlation analysis for climate field reconstructions over Europe. *Journal of Climate* 26 851–867.

Yang, X., Xing, K., Shi, K. and Pan, Q. (2008). Joint state and parameter estimation in particle filtering and stochastic optimization. *Journal of Control Theory and
Applications 6 215–220.

Zhang, Z., Mann, M. E. and Cook, E. R. (2004). Alternative methods of proxy-based climate field reconstruction: application to summer drought over the conterminous United States back to AD 1700 from tree-ring data. The Holocene 14 502–516.