*Submitted to the Annals of Applied Statistics*
arXiv: arXiv:0000.0000

# RECONSTRUCTING PAST TEMPERATURES 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, and explore the effects of including external climate forcings within the reconstruction and of accounting for short-memory and long-memory features. Our reconstructions are 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 is conducted through systematic comparisons between reconstructions that assume no memory; short memory autoregressive models; and long memory fractional Gaussian noise models.

We use Bayesian estimation to fit the model parameters and to perform separate reconstructions of land-only and combined land-and-marine temperature anomalies. For model formulations that include forcings, both exploratory and Bayesian data analysis provide evidence against models with no memory. Model assessments indicate that models with no memory underestimate uncertainty. However, no single line of evidence is sufficient to favor short memory models over long memory ones, or to favor the opposite choice. When forcings are not included, the long-memory models appear to be necessary. While including external climate forcings substantially improves the reconstruction, accurate reconstructions that exclude these forcings are vital for testing the fidelity of climate models used for future projections.

Finally, we use posterior samples of model parameters to arrive at an estimate of the transient climate response to greenhouse gas forcings of 2.5°C (95% credible interval of [2.16, 2.92]°C), which is on the high end of, but consistent with, the expert-assessment-based uncertainties given in the recent Fifth Assessment Report of the IPCC.

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

*Keywords and phrases:* external forcings, long-memory, proxies, temperature reconstruction

**1. Introduction.** An understanding of recently observed and projected future climate changes (Stocker 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. In addition, reconstructions of past climate allow for important out-of-sample assessments of the Atmosphere–Ocean General Circulation Models (GCM) that are used to project future climate under various emissions scenarios (Masson-Delmotte et al., 2013; Flato et al., 2013). While there is 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: green house gas concentrations, volcanic aerosol concentrations, and variations in solar irradiance. The proxies, in turn, are modeled as linear in the latent temperature process. Motivated by existing evidence of long-range correlation in temperature series (e.g., Brody, Syroka and Zervos, 2002; Benth and Šaltytė-Benth, 2005; Huybers and Curry, 2006; Imbers et al., 2014), we explore the effects of specifying white noise (no memory), autoregressive (short memory), and long-memory correlation structures for the two error processes. To our knowledge, this is the first ensemble-based paleoclimate reconstruction that includes the effects of climate forcings, and the first systematic investigation of error structure in 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 (Bindoff 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.

The conceptual study of Li, Nychka and Ammann (2010) demonstrated that temperature reconstructions are improved when information about the climate forcing is included in the reconstruction. We therefore explore the effects of including these three major external forcings in our reconstructions, reporting results for both cases. While the forcings are expected to improve the reconstructions, we note that reconstructions that exclude the forcings are necessary for the evaluations of GCMs (Masson-Delmotte et al., 2013; Flato et al., 2013), to avoid the circularity of using the same forcings in the simulation of past climate and the reconstruction used to assess the simulation.

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).

The assessment of long memory behavior in hierarchical models is complicated by the fact that graphs of the autocorrelation and partial autocorrelation functions (acf and pacf) are generally not adequate diagnostic tools. In addition, the short data streams we are faced with disallow reliance on known asymptotic properties, while 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.

Motivated by the limitations of the data, and our goal of using a robust model, we focus on a simple long-memory model: linear regression with fractional Gaussian noise (fGn) errors. The theoretical question of estimating memory length for non-self-similar models, such as our hierarchical linear model, 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. Online Supplement A.1 in Barboza et al. (2014) 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 single 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 parameters, 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, including parameter posterior distributions and model validation metrics; it compares models with different error structures and which include or exclude the climate forcings. We also 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 (see Barboza et al. (2014)).

**2. Data sets.** The analysis makes use of three distinct data sources: instrumentally 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 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 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 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

[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.

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 appropriate memory lengths in the error structures of the residuals at both levels.

The Bayesian modeling framework is closely related to stochastic filtering methods. An interesting application of classical 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 not an optimal choice for our exploration of short versus long-memory models in paleoclimate reconstructions. In contrast, 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, dimension 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

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.

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 apply a sequence of steps to reduce the number of proxies while attempting to retain as much climatically useful information as possible.

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. 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). 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 B.1 and Table B.1 in Online Supplement B for details (see Barboza et al. (2014)). The single lacustrine proxy included in the

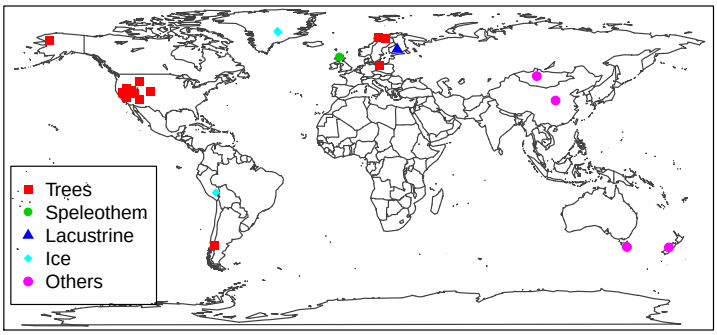



Fig 1: Geographical distribution of the 25 proxy series.

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 heterogenous 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 $\widehat{a_0}, \ldots, \widehat{a_{25}}$ provide a weighted average of proxies that maximizes the explained variance. Denote the reduced proxy as $RP_t$, then

$$RP_t = \widehat{a_0} + \sum_{i=1}^{25} \widehat{a_i} P_{i,t}. \tag{3.1}$$

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. 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 (many of which are coastal) are better at predicting global land and sea temperature than global land-only temperatures. Note that colinearity is not an issue as the $P_{i,t}$ do not feature strong correlations with one another, and in addition our interest lies in the linear combination of $P_{i,t}$ rather than the coefficients $\widehat{a}_i$.

The geophysical distribution of the weights (in percentage of absolute value) is displayed in Tables B.3 and B.4 in Online Supplement B of Barboza et al. (2014). 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 calibrating it, 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 4.3.

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.

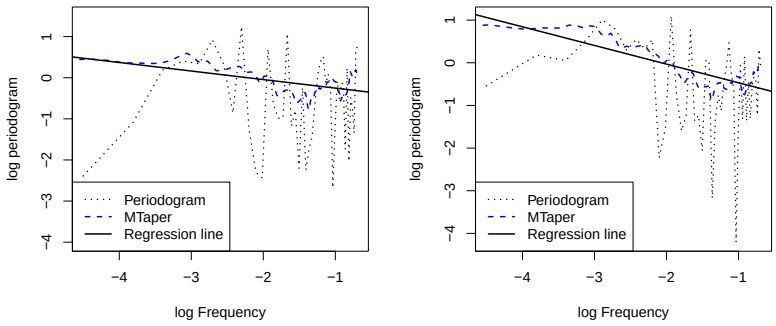
3.2. *Examination of long-memory correlation in the proxy data.* While the temperature-proxy relationship is almost universally assumed to be linear (e.g., Luterbacher et al. (2004), Rutherford et al. (2005), Li, Nychka and Ammann (2010), Tingley and Huybers (2010b), Kaufman et al. (2009), McShane and 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,

$$RP_t = \alpha_0 + \alpha_1 T_t + \sigma_p \eta_t, \tag{3.2}$$

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.

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 $\gamma(n) := \mathrm{cov}\,(X_{t+n}, X_t)$ 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 Online Supplement A in Barboza et al. (2014) for a brief description for each estimator). Figure 2 shows both estimators on a log-log scale for the HAD and CRU datasets, 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




(a) HAD Dataset (b) CRU Dataset

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.

periodogram are less striking than the multitaper estimate, but still show a negative slope in log-log space.

To examine more formally the long memory behavior of the residuals, we employ the test developed by Robinson (1995) (Section 3.4 presents results of alternative tests). To introduce the idea of this method 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. This 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 $\lambda$. 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 Online Supplement A in Barboza et al. (2014). The FARIMA model depends on a parameter usually denoted by

$d = H - 1/2$, and features a spectral density with the same low-frequency and long memory asymptotics as fGn.

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 for the memory length 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.

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. Both datasets therefore show strong evidence, according to Robinson's test, in favor of rejecting the null hypothesis of $H = 0.5$. Note that, the test, while consistent with long memory, does not provide evidence in favor of long-memory correlations over shorter non-zero ones; in the model-comparison exercises below (Section 4.2), we also consider models which contain short-memory, AR(1) errors.

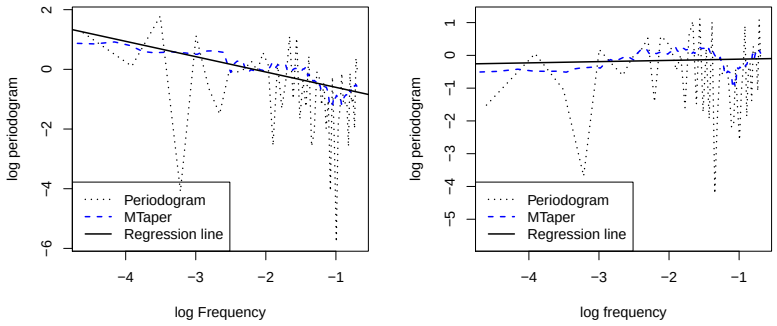
3.3. *Examination of long-memory behavior in the temperature anomalies.* In the specification of the process level of the hierarchical model, we follow Li, Nychka and Ammann (2010) and model the latent temperature as linear in the external forcings. We apply the following transformations to the forcings, 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 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,

$$T_t = \beta_0 + \beta_1 S_t + \beta_2 \tilde{V}_t + \beta_3 \tilde{C}_t + \sigma_T \epsilon_t, \tag{3.3}$$

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,




(a) HAD Dataset (b) CRU Dataset

Fig 3: Spectral estimates in log-log scale. The regression line is computed by regressing the logarithm of multitaper estimator on log-frequencies. The frequency units are cyle/year.

$d = 0.4,\ 0$), with Hurst parameter $H = d + 0.5 = 0.9$. Benth and Šaltytė-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. (2014) 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 the 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 strong evidence against no memory for the HAD data set, but not for the CRU data set.

As there is value in reconstructions that exclude the forcings (e.g., for the purpose of General Circulation Model assessment), we also consider a reduced form of the process-level model that exclude the forcings, and models climate variability as a purely stochastic process. Applying Robinson's test

TABLE 2
*Results of Beran's test applied to the residuals from the HAD and CRU datasets, for both the proxy [Eq. 3.2] and instrumental [Eq. 3.3] equations, with or without forcings, under three null hypotheses*

| **Model** | **fGn** | **AR(1)** | **AR(2)** |
|---|---|---|---|
| HAD-Proxy | 0.77 | 0.40 | 0.76 |
| CRU-Proxy | 0.91 | 0.72 | 0.92 |
| HAD-Temp | 0.56 | 0.58 | 0.59 |
| CRU-Temp | 0.73 | 0.33 | 0.40 |
| HAD-Temp (*No-Forcings*) | 0.61 | 0.63 | 0.67 |
| CRU-Temp (*No-Forcings*) | 0.46 | 0.19 | 0.47 |

to the CRU and HAD data sets results in p-values of $2.12 \times 10^{-10}$ for both cases, where we can note that the amount of evidence against no-memory increases when we exclude the forcings.

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 a particular stochastic process model (e.g. fGn) to a realization of a time series. Let $X_t$ be a stationary Gaussian process with spectral density $f(\lambda)$, whose realization one observes. When testing for fGn for instance, 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 certain changes of scale (see Online Supplement A in Barboza et al. (2014) for additional technical details).

We performed Beran's test on six datasets: the four residuals from the HAD and CRU datasets, for both the proxy [Eq. 3.2] and instrumental [Eq. 3.3] equations, and the two HAD and CRU temperature data series themselves with no forcings. To test the presence of memory, we use three distinct memory structures: fractional Gaussian noise, AR(1), and AR(2). The null hypothesis in each test is that the data comes from a spectral density equal to that of the given memory structure. Thus a non-rejection of the null is not inconsistent with the tested memory structure. For our eighteen Beran's tests, the corresponding p-values are shown in Table 2. The results indicate that Beran's test cannot reject the null in any of the eighteen cases; this is consistent with the presence of memory, but the tests do not point to a preferred memory structure.

Finally, we apply the test proposed by Davies and Harte (1987); see section A.1.2 for technical details. The fGn is used as the underlying parametric model for this test, and the null and alternative hypotheses are identical to

TABLE 3
*Results of Davies and Harte's test applied to the residuals from the HAD and CRU datasets, for both the proxy [Eq. 3.2] and instrumental [Eq. 3.3] equations, with or without forcings, under the null hypothesis of no memory*

| Model | Davies & Harte |
|---|---|
| HAD-Proxy | 0.046 |
| CRU-Proxy | 0.000 |
| HAD-Temp | 0.010 |
| CRU-Temp | 0.436 |
| HAD-Temp (*No-Forcings*) | 0.000 |
| CRU-Temp (*No-Forcings*) | 0.000 |

Robinson's test: $H = 0.5$ (no memory) versus $H > 0.5$ (long-memory). Thus, in contrast with Beran's test, rejection of the null is evidence against no memory. As in Beran's test, we use the four residuals from the HAD and CRU datasets, for both the proxy [Eq. 3.2] and instrumental [Eq. 3.3] equations, and the two HAD and CRU series with no forcings. P-values in Table 3 show that the null can be rejected in three out of four cases when we include forcings within the models, and in the two cases without forcings. In fact, the evidence against no-memory increases when we exclude forcings.

No single method employed here is a perfect indicator for the presence or absence of memory in our error 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 memory, long or short, cannot be ignored in developing models for the residuals or for the HAD and CRU series themselves. In Section 4, we further investigate the memory properties of the residual processes, via Bayesian parameter estimates and reconstruction validation measures.

3.5. *Hierarchical Bayesian model with long or short-memory errors.* Given the statistical evidence for long or short memory correlation in the empirical residuals from Eqs (3.2) and (3.3), and the implication for fGn or AR model by Beran's test, we explore the results of modeling the errors using either fGn or AR processes. As the strategy for fitting the hierarchical Bayesian reconstruction is similar in each case, we present details for the more computationally involved fGn error assumption. Comparisons between various modeling choices (long memory vs. short memory vs. no memory; with or without forcings) are given in Section 4.3. A summary of the data and process levels of the hierarchical model is as follows:

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

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 $\boldsymbol{\alpha} := (\alpha_0, \alpha_1)^{\mathrm{T}}$, $\boldsymbol{\beta} := (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, $\sigma_1^2$, $\sigma_2^2$, $H$ and $K$:

- $\boldsymbol{\alpha} \sim N((0,1)^{\mathrm{T}}, \mathbf{I}_2)$; $\boldsymbol{\beta} \sim N((0,1,1,1)^{\mathrm{T}}, \mathbf{I}_4)$;
- $\sigma_T^2 \sim IG(2; 0.1)$, $\sigma_P^2 \sim IG(2; 0.1)$;
- $H \sim \mathrm{Unif}(0,1)$; $K \sim \mathrm{Unif}(0,1)$.

where $\mathbf{I}_n$ is the identity matrix of dimension $n$.

Let $\mathbf{T}_u = (T_{1000}, \ldots, T_{1899})$ denote the vector of unknown temperatures and $\mathbf{T}_0 = (T_{1900}, \ldots, T_{1998})$ the vector of instrumental temperatures. Our goal is to infer $\mathbf{T}_u$ based on $\mathbf{T}_0$, $RP$, $S$, $\tilde{V}$ and $\tilde{C}$. The full conditional posterior distributions of $\mathbf{T}_u$ 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 Online Supplement A in Barboza et al. (2014). 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).

**4. Numerical results.** The diagnostic tests in Section 3, while providing no conclusive evidence for the presence of long or short memory, indicate the possibility of certain correlations. In order to further investigate appro-

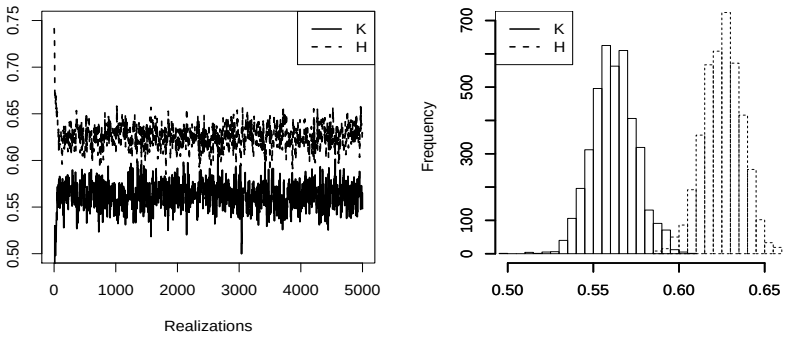
priate models for error structures and to assess the benefit of incorporating external forcings in the reconstruction, we compare eight model variants on the basis of their parameter estimates and reconstruction validation metrics:

A : Possible long memory ($H$ and $K$ not fixed), with external forcings.
B : Possible long-memory error in (3.2) and AR(1) error in (3.3), with external forcings.
C : AR(1) error in (3.2) and possible long-memory error in (3.3), with external forcings.
D : AR(1) errors in (3.4), with external forcings.
E : No memory ($H = K = \frac{1}{2}$), with external forcings.
F : Possible long memory ($H$ and $K$ not fixed), no external forcings ($\beta_i = 0, i = 1, 2, 3$).
G : AR(1) errors in (3.4), no external forcings ($\beta_i = 0, i = 1, 2, 3$).
H : No memory ($H = K = \frac{1}{2}$), no external forcings ($\beta_i = 0, i = 1, 2, 3$).

The AR(1) model is included as it features short memory – an intermediate model between assuming fGn and assuming uncorrelated white noise. We refer to scenarios E and H as having no memory, as they are based on Gaussian white noise errors that are independent and thus have no memory. Scenario B allows for a long-memory model for the proxies while assuming short-memory in the temperature residuals, while scenario C reverses the assumptions of scenario B.

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 Online Supplement B (see Barboza et al. (2014)). 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.* We first examine parameters estimates using the HAD data set and including the forcings. Figure 4 shows trace plots and histograms of the $H$ and $K$ parameters that are responsible for long memory in scenario A. 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. 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



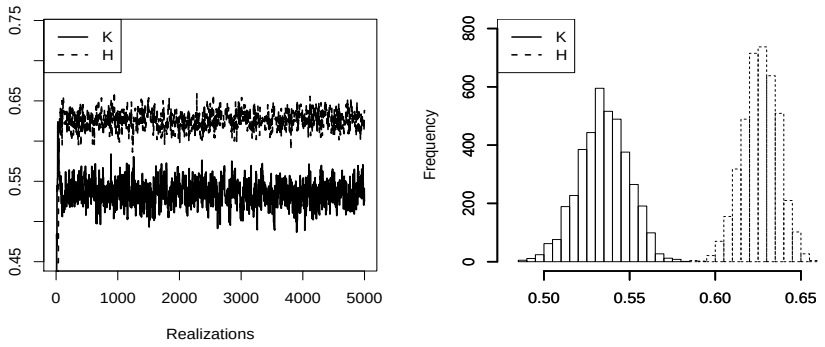

(a) Traceplots (b) Histograms

Fig 4: Bayesian estimation of $H$ and $K$ based on HAD Dataset, Scenario A.

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 resulting in a longer timescale response to changes in the forcings.

For Scenarios B, C and D, both HAD and CRU show that all AR(1) parameter estimates are significantly greater than zero, and all long-memory parameter estimates are significantly greater than 0.5 (See figures B.2, B.3 and B.4 in the Online Supplement B of Barboza et al. (2014)). For Scenarios F and G, which exclude the forcings, and have respectively long and short memory, Bayesian posteriors for the memory parameters provide evidence against models with no memory at higher levels of certainty than for models that include forcings, especially in the CRU case; see Figure B.5 in Online Supplement B (see Barboza et al. (2014)). These results indicate that while there is a certain amount of memory in the error structures, there is insufficient evidence to select between short or long memory assumptions. In the subsequent section, we resort to reconstruction validation metrics to compare different models.

Posterior samples for the process-level regression coefficients (the $\beta_i$) for Scenario A show that the transformed volcanic and greenhouse gas forcing series are meaningful predictors of the temperature evolution for both HAD and CRU, while solar irradiance is less influential (Figures B.8 and B.12). While the forcings are useful predictors of past temperatures, we stress that the reconstruction that exclude the forcings are also of scientific interest.




(a) Traceplot (b) Histogram

Fig 5: Bayesian estimation of $H$ and $K$ based on CRU Dataset, Scenario A.

Such reconstructions may not provide the most accurate estimates of past climate fluctuations, but provide necessary test beds for assessing the GCMs used to project future climate, since comparisons between forcings-based reconstructions, and GCM simulations which are based on the same forcings, would pose circularity issues.

4.2. *Validation measures.* We provide quantitative assessments of the eight 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); root mean squared error (RMSE); 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 Probability Score (CRPS). The ECP measures the accuracy of the uncertainty quantification and values closer to nominal level are more desirable, while the IS and CRPS provide more nuanced assessments of the posterior predictive distributions, rewarding both calibration and sharpness simultaneously; details of these scoring rules are available in Gneiting and Raftery (2007); Gneiting, Balabdaoui and Raftery (2007); Gschlößl and Czado (2007), and Online Supplement A.3 in Barboza et al. (2014). For convenience, we report the negative IS and CRPS so that smaller values indicate higher quality predictions.

Table 4 summarizes the quantitative assessments of the reconstructions

TABLE 4

*Validation measures for the eight reconstruction scenarios, using both HAD and CRU datasets. Scenarios F, G, and H, which include no forcings, are italicized in this table.*

| | Scenarios | Sq. Bias | Variance | RMSE | $ECP_{95}$ | $ECP_{80}$ | $IS_{95}$ | $IS_{80}$ | CRPS |
|---|---|---|---|---|---|---|---|---|---|
| HAD | A | 0.016 | 0.012 | 0.168 | 92.9 | 74.7 | 0.062 | 0.178 | 0.208 |
| | B | 0.017 | 0.013 | 0.171 | 92.9 | 74.7 | 0.062 | 0.179 | 0.205 |
| | C | 0.015 | 0.011 | 0.160 | 90.9 | 72.7 | 0.064 | 0.179 | 0.212 |
| | D | 0.015 | 0.011 | 0.162 | 90.9 | 74.7 | 0.063 | 0.176 | 0.209 |
| | E | 0.014 | 0.010 | 0.154 | 90.9 | 69.7 | 0.060 | 0.171 | 0.195 |
| | *F* | *0.055* | *0.072* | *0.356* | *99.0* | *84.8* | *0.110* | *0.323* | *0.229* |
| | *G* | *0.081* | *0.071* | *0.390* | *94.9* | *75.8* | *0.118* | *0.389* | *0.259* |
| | *H* | *0.113* | *0.059* | *0.415* | *82.8* | *59.6* | *0.168* | *0.511* | *0.304* |
| CRU | A | 0.032 | 0.025 | 0.238 | 91.9 | 73.7 | 0.084 | 0.251 | 0.245 |
| | B | 0.031 | 0.025 | 0.235 | 91.9 | 75.8 | 0.081 | 0.245 | 0.237 |
| | C | 0.033 | 0.024 | 0.238 | 91.9 | 71.7 | 0.090 | 0.258 | 0.252 |
| | D | 0.032 | 0.023 | 0.234 | 90.9 | 70.7 | 0.087 | 0.255 | 0.245 |
| | E | 0.031 | 0.024 | 0.235 | 91.9 | 73.7 | 0.085 | 0.250 | 0.242 |
| | *F* | *0.089* | *0.097* | *0.432* | *97.0* | *78.8* | *0.131* | *0.416* | *0.274* |
| | *G* | *0.120* | *0.095* | *0.464* | *90.9* | *75.8* | *0.150* | *0.482* | *0.303* |
| | *H* | *0.148* | *0.080* | *0.477* | *84.8* | *62.6* | *0.206* | *0.570* | *0.335* |

RMSE: Root Mean Square error, $ECP_\beta$: Empirical Coverage Probability at $\beta$% confidence level, $IS_\beta$: Interval Score at $\beta$% confidence level, CPRS: Continuous Ranked Probability Score.
* HAD and CRU refer to the two instrumental data sets, with HAD including the oceans.

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, as well as the IS and CRPS (compare Scenario A to F, Scenario D to G, and Scenario E to H). This corroborates the fact that the posterior distributions of the coefficients for both the volcanic and green house gas forcing series are significant. Moreover, the widths of the 95% credible intervals are likewise narrower when the external forcings are included (see Figure 8, below, and Figure B.16).

When external forcings are included in the reconstruction, the squared biases, variances and RMSEs are generally similar across the different error models for each of the two data sets. For the HAD data set, and amongst reconstructions that include forcings, Scenarios A and B are optimal in terms of ECP; Scenario E in terms of CRPS; and there are no appreciable differences in IS. Note that Scenario E exhibits the worst ECP, indicating an underestimation of uncertainty compared to Scenarios A and B. This is consistent with the rejection of no-memory models in our tests in Section 3. For the CRU data set, Scenario B is optimal in terms of ECP and CRPS, and again there is no appreciable difference in terms of IS. Based on these validation measures, while there continues to be support for memory models, there is no clear indication of a single, best model for the error structures amongst

the reconstructions that include forcings, with Scenarios A and B featuring comparable performance metrics. Indeed, tests for selecting between long and short memory models for climate time series are often inconclusive (e.g. Percival, Overland, and Mofjeld, 2001).

When forcings are not included, the greater variability of validation metrics across the scenarios allows for more meaningful ranking of the error correlation assumptions. For both data sets, Scenario F is optimal in terms of squared bias, RMSE, IS and CRPS. For the HAD data set, Scenario G is optimal in terms of ECP at the 95% level but is equally distant from the nominal level as Scenario F at the 80% level, while for the CRU data set, ECP favors Scenario F. In general, the results indicate that when forcings are not included the long-memory models play an important role in capturing the correlation structure in proxies and temperature and should be employed in the hierarchical model. As discussed in Li, Nychka and Ammann (2010), reconstructions are improved when information is included at a broad range of frequency scales. In the absence of forcings, which feature long-range correlations and low-frequency behavior, the inclusion of more highly structured noise processes leads to marked improvements in the reconstructions.

4.3. *Temperature reconstruction results.* According to validation measures in Table 4, the reconstruction scenarios that include forcings are similar to one another. Here we focus on Scenario B due to slightly better validation measures for both HAD and CRU datasets. Figure 6 shows the Scenario B temperature reconstruction together with 95% point-wise credible intervals, using the HAD dataset. 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 Fig. B.14), is qualitatively similar, but features higher variance due to the more variable CRU temperatures.

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 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 in Scenarios A, B, E, F and H, as well as the actual HAD observations. The Scenarios that include forcings (A, B, E) result in reconstructions that are qualitatively similar to one another and feature good qualitative agreement with the observations, with Scenario E exhibiting slightly narrower credible intervals.

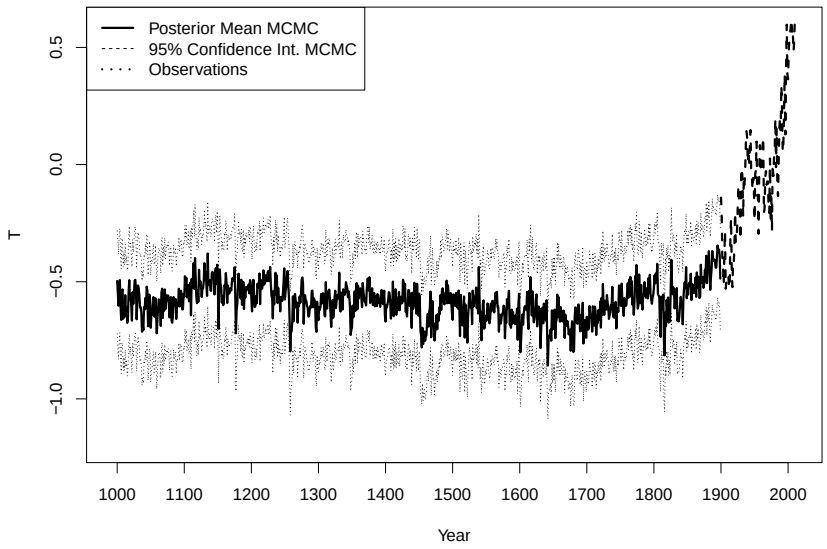



Fig 6: Temperature reconstruction (1000-1899) using the HAD dataset, Scenario B.

Reconstructions resulting from scenarios that exclude the forcings (F and H) feature greater divergence from the observations – particularly for Scenario H, which models the error structure as white noise. Results are similar for the CRU data set (see Figure B.15). 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). Repeating the reconstructions with the single lacustrine record excluded from the reduced proxy leads to similar results; see Fig. B.17.

4.4. *MCMC Diagnostics.* To establish convergence of the MCMC samples, we examine trace plots (Figures B.6–B.13), 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. We present diagnostic results for Scenario A as it represents the most complex model for estimation. If the

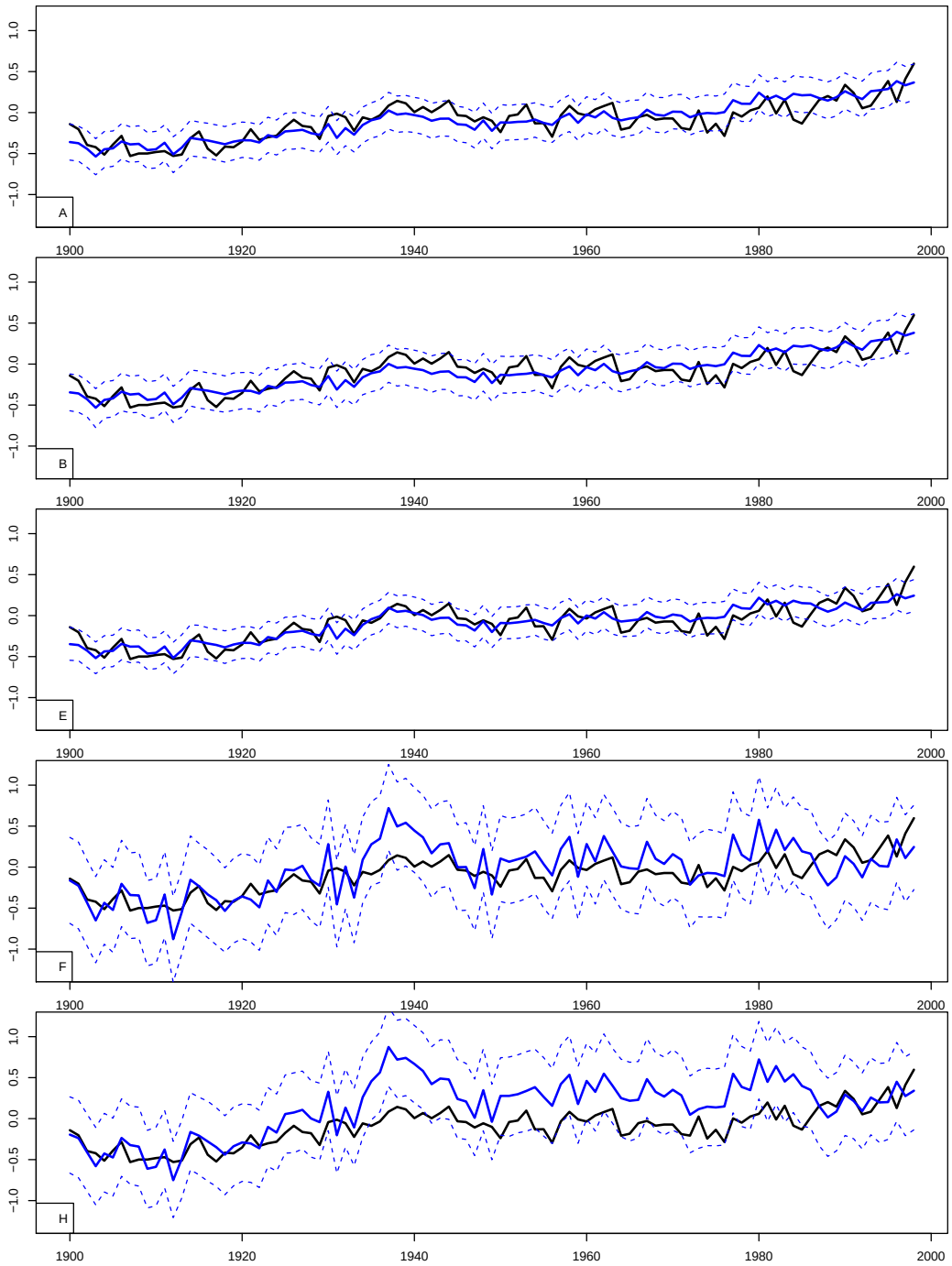



Fig 7: Temperature reconstruction (1900-1998) using the HAD dataset under Scenarios A,B,E,F and H. Black: Observations; Dashed blue: 95% credible intervals MCMC; Solid blue: posterior mean MCMC.

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 `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.

TABLE 5
*Individual and multivariate potential scale reduction factors (PSRF) with the 95% upper bounds (UB) for individual PSRFs.*

| | $\alpha_0$ | $\alpha_1$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma_P^2$ | $\sigma_T^2$ | $H$ | $K$ | Mul. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HAD PSRF | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.00 | 1.01 | 1.00 | 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 | 1.01 | 1.00 | 1.01 | 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) assume no memory in the error processes, do not include 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 online[2] as decadally smoothed time series, as Mann et al. (2008a)

[2] http://www.ncdc.noaa.gov/paleo/pubs/mann2008/mann2008.html

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[3] with cutoff frequency equal to 0.1 cycles/year.

Figure 8 compares our reconstructions using the HAD data and under Scenarios A, B, E, F, and H, 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).

The model settings of Mann et al. (2008a) are most similar to our Scenario H, 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 H 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 H results. Results are similar when using the CRU data set (Figure B.16).

To facilitate numerical comparisons with the Mann et al. (2008a) reconstruction, we re-calculate validation metrics for Scenario H 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 H results and the Mann et al. (2008a) results is in terms of squared bias, with 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 H, on par with results from our annually resolved Scenarios A, B, C, D and E. 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 H is smaller than that in Mann et al. (2008a), but due to the relatively large bias, the ECPs

[3]Our calculations are based on the Matlab code associated with Mann et al. (2008a), posted online at `http://www.ncdc.noaa.gov/paleo/pubs/mann2008/mann2008.html`. We smooth using the `filtfilt` command in the R package "signal".

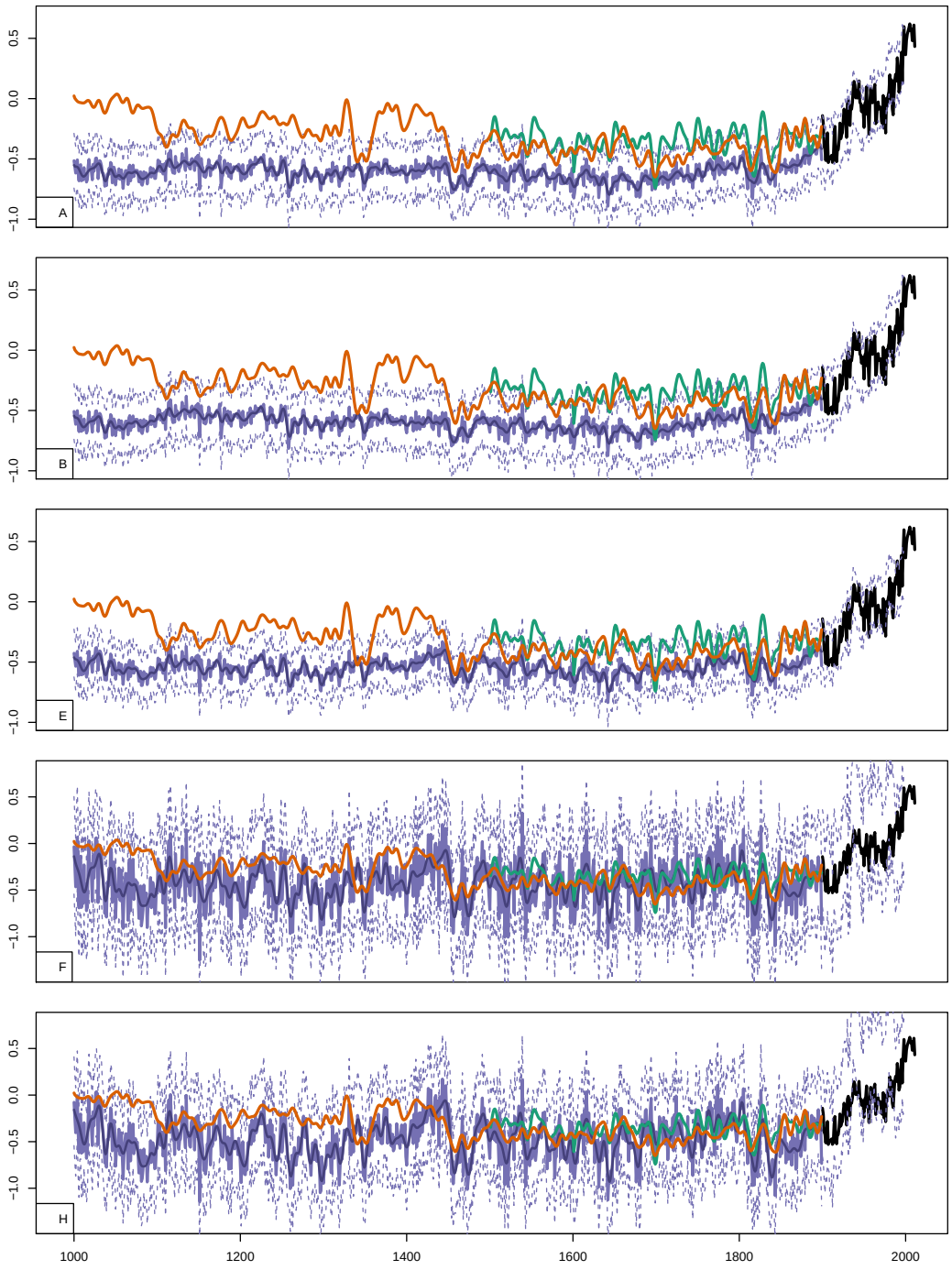



Fig 8: Comparisons between Scenarios A, B, E, F, H and CPS and EIV reconstructions in Mann et al. (2008) using the HAD dataset. Black: Observations; Purple: posterior mean reconstruction with 95% credible intervals; Orange: EIV; Green: CPS. Dark Purple line: mean of smoothed posteriors.

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 H, and like the RMSE, are similar to those for our annually resolved Scenarios A, B, C, D and E which carry small squared bias by including the forcings (see Table 4).

TABLE 6
*Comparison between Scenarios H and CPS and EIV Reconstructions in Mann et al. (2008)*[*]

| | Scenarios | Sq. Bias | Variance | RMSE | $ECP_{95}$ | $ECP_{80}$ | $IS_{95}$ | $IS_{80}$ |
|---|---|---|---|---|---|---|---|---|
| HAD | H (smoothed) | .100 | .012 | .335 | 41.4 | 33.3 | 0.46 | 0.71 |
| | CPS | .009 | .024 | .183 | 100.0 | 96.9 | 0.06 | 0.16 |
| | EIV | .003 | .022 | .157 | 99.0 | 99.0 | 0.06 | 0.23 |
| CRU | H (smoothed) | .121 | .016 | .371 | 48.5 | 36.4 | 0.45 | 0.73 |
| | CPS | .017 | .025 | .207 | 99.0 | 99.0 | 0.07 | 0.25 |
| | EIV | .006 | .021 | .163 | 98.0 | 98.0 | 0.07 | 0.17 |

[*]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 $M_t \pm \frac{1.96}{2}S$, where $M_t$ is their predicted temperature mean.

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 H, as numerous lines of evidence indicate that Scenario H is the least appropriate of the eight scenarios explored here: validation metrics and scores (Table 4) are generally worst for Scenario H; the inclusion of the forcings is motivated by the scientific understanding of their connection with temperatures; and the inclusion of the long memory processes in the absence of forcings is driven by the structure of the data. Indeed, we view Scenario H 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 Scenarios A, B, C, and D, which include the forcings as well as short and/or long memory processes, are comparable in terms of RMSE and Interval Scores to the decadally resolved Mann et al. (2008a) results while featuring ECPs which are closer to their nominal values.

Finally, we note that the proxy selection and modeling treatments do differ between our Scenario H 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, 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 smoothed Scenario H and those for the Mann et al. (2008a) reconstructions are not surprising.

5.2. *Transient climate response.* 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 functionally related to the parameter $\beta_3$, and the resulting estimates of TCR are based on the instrumental temperature record 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:

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

where $\sigma(\log \mathbf{C})$ is the standard deviation of the logarithm of the GHG series $\mathbf{C}$, 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 which we can compute a non-parametric estimator of the probability density function for TCR that accounts for the uncertainties in all other parameters in the model.

We present results of TCR estimates using the global land and marine HAD data set, for the five scenarios that include the forcings: Scenarios A, B, C, D, and E (Figure 9). There is substantial variability between the TCR estimates from the five scenarios. TCR estimates are the lowest and most sharply peaked for the memory-free Scenario E, with a median around 2.39°, and an approximate 95% credible interval of $[2.16, 2.63]$°C. The TCR distributions become progressively broader as more memory is included, in

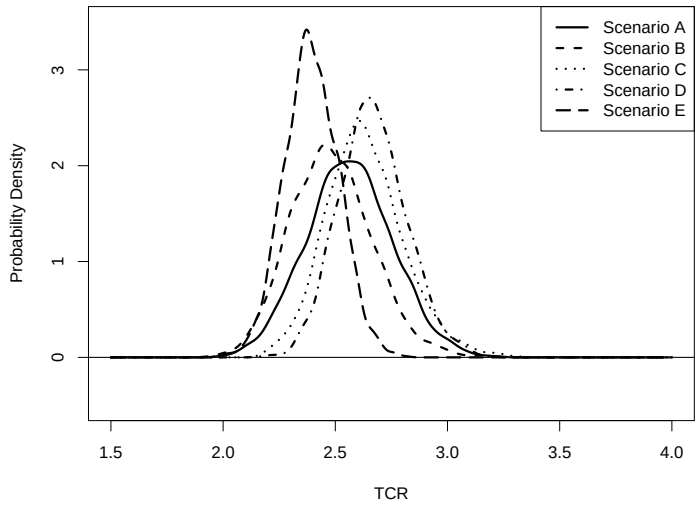



Fig 9: Estimates of probability density function for Transient Climate Response (TCR) in degree Celsius for Scenarios A, B, C, D and E (HAD).

scenarios D, then in B and C, and finally in the fully long-memory Scenario A which features the broadest 95% credible interval of $[2.19, 2.95]$°C. A quantitative explanation for this increasing uncertainty behavior can be found by inspecting the formula for the covariance matrix $\Omega_\beta$ of the posterior distribution of the vector $\beta$ given $T$: from formula (A.4) in the Online Supplement A (see Barboza et al. (2014)), one sees that $\Omega_\beta$ is the inverse of a matrix which is affine in $\Sigma_K^{-1}$, i.e. affine in the inverse of the covariance matrix for the noise model being used in each scenario. It is known (see for instance Palma and Bondon (2003)[4] that the magnitude (e.g. the operator norm) of $\Sigma_K$ increases with memory length; this and the formula for $\Omega_\beta$ can explain the increasing behavior we observe in Figure 9.

On the other hand, the progression of posterior medians for the TCRs is not monotone. Scenario D (AR(1) errors) features the largest median

[4]In this paper, the authors provide the estimate $\lambda_{n,n} \asymp n^{2H-1}$ for the top eigenvalue of the covariance matrix of a vector of $n$ contiguous terms of a stationary sequence whose covariance matrix $\rho$ satisfies $\rho(n) \sim cn^{2H-2}$, which is the case for our fGn sequence. Thus indeed $\lambda_{n,n}$ is roughly increasing in $H$ for all $H \in (0.5, 1)$. Palma and Bondon (2003) state this result in the case of the ARFIMA process, in Example 2 on pages 99-100, but an inspection of their proof shows that the result holds for all $\rho$ satisfying the above asymptotics.

TCR value: 2.66°C, followed by Scenario C with short memory in the proxy model: 2.62°C. The two scenarios with long memory in the proxy model, A and B, have lower median TCR values respectively 2.56°C and 2.47°C, with Scenario E (no memory) having the lowest median of 2.39°C, as reported above. The formula for the posterior distribution of the vector $\beta$ given $T$ is again helpful: the posterior mean is the product of the increasing $\Omega_\beta$, as discussed above, and of a matrix $\Delta_\beta$ which is affine in $\Sigma_K^{-1}$, thus with presumably decreasing magnitude with respect to memory length; the competition between these two effects could induce non-monotonicity with respect to memory length.

To arrive at a best estimate of TCR, we mix with equal weights the posterior estimates from Scenarios A and B, yielding a median TCR of about 2.5°C with a combined 95% credible interval of about $[2.16, 2.92]$°C. Scenarios A and B feature superior validation metrics as compared with Scenarios C and D, while Scenario E presumably under-reports TCR uncertainty because it is based on a model that lacks memory. We therefore focus on TCR estimates derived from an equally weighted mixture of estimates from scenarios A and B. It is not possible to select between these two scenarios using model diagnostics, and, as both include memory, our estimates are conservative with respect to uncertainty.

It is instructive to compare our five TCR distributions reported here with the consensus (expert assessment) recently released in the IPCC's Fifth Assessment Report (Bindoff et al., 2013; Collins et al., 2013), where TRC is reported as "likely" within the interval $[1, 2.5]$°C and "extremely unlikely" to exceed 3.0°C. Using the IPCC's definitions/guidance on uncertainty language, these expert assessments can be interpreted as meaning that the probability of the estimated TCR from one of our scenarios falling within the $[1, 2.5]$°C interval should exceed 0.66, while the probability that the estimated TCR exceeds 3.0°C should not exceed 0.05. For all scenarios reported in Figure 9, the posterior probability that the TCR exceeds 3.0°C is in all cases lower than 0.05. This exceedance probability is essentially zero for Scenario E, which features the narrowest TCR distribution: scenario E presumably under-estimates uncertainty by using no memory for modeling errors. As for falling in the interval $[1, 2.5]$°C with probability around 0.66, Scenario E does satisfy this condition; Scenario B nearly does; for Scenario A the probability is closer to 0.5; but Scenarios C and D, with their significantly higher median values, fail the condition by some margin. Our best estimate, derived from mixing scenarios A and B, meets the TCR upper bound condition: the probability that it exceeds 3.0°C is about 0.011. It falls slightly short of meeting the confidence interval condition: the proba-

bility that it falls within the interval $[1, 2.5]^\circ$C is about 0.47.

All of our reported TCRs are on the high side compared to the latest IPCC consensus, and as compared with several specific recent studies which have arrived at TCR estimates by combining information from models and the instrumental temperature record. Gillett et al. (2012) produce a TCR estimate of $1.3 - 1.8^\circ$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, $0.9 - 2.3^\circ$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 $0.7 - 2.5^\circ$C, but caution against strong conclusions based on a such a short time interval. A particularly high estimate of TCR, of at least $2.5 - 3.6^\circ$C, is reported by Tung et al. (2008), based on an analysis of the 11-year solar cycle.

Hence both the specific model-data fusion studies discussed in the previous paragraph and in Bindoff et al. (2013), as well as the synthesis provided by the IPCC Fifth Assessment report, generally feature broader uncertainties and are peaked at lower values as compared our posterior estimates of TCR. Indeed, only one of the estimated TCR distributions shown in Figure 10.20 of Bindoff et al. (2013) is peaked at a value greater than $2^\circ$C, while the high estimate of (Tung et al., 2008) is explained as resulting from solar forcing having a different mechanistic effect on climate (Bindoff et al., 2013). Interestingly, the single plotted TCR distribution peaked at greater than $2^\circ$C is that of Harris et al. (2013), which estimates TCR using a Bayesian approach that combine information from GCMs and recently observed temperature changes. A possible cause for the narrower uncertainties and higher TCR values estimated here is the more extensive use of data, in terms of both variety (instrumental temperatures, proxies, and estimates of $CO_2$, volcanic, and solar forcings) and duration (observations over the last millennium).

**6. Conclusions and Discussion.** We use a comprehensive multiproxy data set to produce new reconstructions of NH temperature anomaly time series back to 1000 AD, and systematically evaluate the effects of including or excluding external drivers of climate variability, and of assuming the error processes feature long, short, or no memory, by considering eight modeling scenarios. Hierarchical Bayesian models are used throughout as they provide a natural framework for integrating the different information sources – proxy and instrumental temperatures observations, and time series of solar,

greenhouse gas, and volcanic forcings. Bayesian inference additionally permits for estimation of all unknown quantities, including past temperatures, and facilitates uncertainty propagation.

While the possibility of long-memory was suggested by exploratory data analysis, and the significance of long memory parameters verified by Bayesian estimation, model diagnostics indicated that short and long memory models yield comparable results provided the climate forcings are incorporated into the reconstructions. 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 estimates are near the upper bound of the expert-derived "extremely likely" interval provided in the IPCC Fifth Assessment Report (Bindoff et al., 2013), they do not violate this uncertainty consensus, and we note that our estimate is based on both the instrumental and paleoclimate records, and does not rely on GCMs.

If the forcings are excluded from the reconstruction, as is necessary for reconstructions to be suitable for GCM assessment exercises, the long-memory processes substantially improve the quality of the reconstructions. The scenario with neither forcings nor memory is 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 indicate 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, 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.

## SUPPLEMENTARY MATERIAL

**Supplement to: "Reconstructing Past Temperatures from Natural Proxies and Estimated Climate Forcings using Short- and Long-Memory Models"**

(doi: ??????????; .pdf). We provide a background on long-memory models, the multitaper estimator and scoring rules together with some calculations of our model's posterior distributions. Finally, we include additional plots and tables.

## References.


Ammann, C. M., Joos, F., Schimel, 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.

Barboza, L., Li, B., Tingley, M. P. and Viens, F. G. (2014). Supplement to: "Reconstructing Past Temperatures from Natural Proxies and Estimated Climate Forcings using Short- and Long-Memory Models" DOI:??????????.

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.

Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I. 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* (T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley, eds.) Cambridge University Press, Cambridge, UK.

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.

Brooks, S. P. and Roberts, G. O. (1997). Assessing convergence of Markov Chain Monte Carlo algorithms. *Statistics and Computing* **8** 319–335.

Christensen, J. H., Hewitson, 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.

Collins, M. R., Knutti, R., Arblaster, J., Dufresne, J.–L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., Wehner, M. and (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* (T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley, eds.) Cambridge University Press, Cambridge, UK.

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.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. and Rummukainen, M. (2013). Evaluation of Climate Models. 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* (T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley, eds.) Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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.

Harris, G. R., Sexton, D .M. H., Booth, B. B. B., Collins, M. and Murphy, J. M., (2013). Probabilistic projections of transient climate change. *Climate Dynamics* **40** 2937–2972.

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. Averyt, 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. (2014). Sensitivity of climate change detection and attribution to the characterization of internal variability. *Journal of Climate* **27** 3477–3491.

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., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Jansen, 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.DCSupplemental/0805721105SI.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.

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., Rouco, J. F. G., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X. and Timmermann, A. (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* (T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley, eds.) Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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.

Palma, W. and Bondon, P. (2003). On the eigenstructure of generalized fractional processes. *Statistics & Probability Letters* **65** 93–101.

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.

Stocker, T. F., Dahe, Q., Plattner, G.-K., 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 Kanikicharla, 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* (T. F. Stocker, Q. Dahe, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley, eds.) Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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.

Tung, K. K. Zhou, J. S. and Camp, C. D. (2008). Constraining model transient climate response using independent observations of solar-cycle forcing and response.. *Geophysical Research Letters* **35** L17707.

Wahl, E. R. and Smerdon, J. E. (2012). Comparative performance 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.

Centro de Investigación en Matemática
Pura y Aplicada (CIMPA)
Universidad de Costa Rica
San José, Costa Rica
E-mail: luisalberto.barboza@ucr.ac.cr

Department of Statistics
Purdue University
West Lafayette, IN
E-mail: viens@purdue.edu

Department of Meteorology
and Department of Statistics
Pennsylvania State University
State College, PA
E-mail: mpt14@psu.edu

Department of Earth and Planetary Sciences
Harvard University
Cambridge, MA
E-mail: mpt14@psu.edu

Department of Mathematics
Purdue University
West Lafayette, IN
E-mail: viens@purdue.edu

Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL
E-mail: libo@illinois.edu