Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

The Value of Multiproxy Reconstruction of Past Climate

Bo LI, Douglas W. NYCHKA, and Caspar M. AMMANN

Understanding the dynamics of climate change in its full richness requires the knowledge of long temperature time series. Although long-term, widely distributed temperature observations are not available, there are other forms of data, known as climate proxies, that can have a statistical relationship with temperatures and have been used to infer temperatures in the past before direct measurements. We propose a Bayesian hierarchical model to reconstruct past temperatures that integrates information from different sources, such as proxies with different temporal resolution and forcings acting as the external drivers of large scale temperature evolution. Additionally, this method allows us to quantify the uncertainty of the reconstruction in a rigorous manner. The reconstruction method is assessed, using a global climate model as the true climate system and with synthetic proxy data derived from the simulation. The target is to reconstruct Northern Hemisphere temperature from proxies that mimic the sampling and errors from tree ring measurements, pollen indices, and borehole temperatures. The forcing series used as covariates are solar irradiance, volcanic aerosols, and greenhouse gas concentrations. The Bayesian model was successful in integrating these different sources of information in creating a coherent reconstruction. Within the context of this numerical testbed, a statistical process model that includes the external forcings can improve the quality of a hemispheric reconstruction when long time scale proxy information is not available. This article has supplementary material online.

KEY WORDS: Bayesian hierarchical model; Forcings; Global climate model; Past temperature reconstruction; Proxies.

1. INTRODUCTION

Understanding the complex dynamics of the Earth's climate system is a grand scientific challenge in the geosciences. Recently this problem has assumed additional importance with the scientific consensus that human activities are responsible for the global warming observed over the last decades (IPCC 2007), and also with the expectation that greenhouse gas concentrations will continue to increase in the 21st century to levels that have not been present for millions of years. Not only will these dramatically change the global climate, but even larger impacts are expected at a regional scale. Part of the difficulty in making projections of climate for 50 or 100 years in the future is that the behavior of the Earth system over such time scales has not been well observed. In our work we wish to consider the climate over the past 1000 years as a baseline to study long-term variability and responses to variations in external *forcings* such as solar radiation, greenhouse gases, atmospheric aerosols (fine particles) and other factors. Also, time spans of this length or longer are useful for testing whether advanced climate models, also known as atmosphere/ocean general circulation models (AOGCM), are able to simulate longer term features of the climate system accurately. AOGCMs are the main scientific tool for estimating future climate in response to different scenarios of greenhouse gas concentrations and so their development and validation are important for quantifying climate change over the next century. A statistical approach for validation is especially useful because the discrepancy between the model simulations and observations can be contrasted with the uncertainty in the observations themselves.

Estimates of the Northern Hemisphere (NH) temperature based on a spatial network of stations taking direct temperature measurements only exists for the past 150 years. Prior to this period the observations are too sparse to provide an adequate estimate of this hemispheric average. Moreover, measurements of the ocean and upper atmosphere are even more restricted in time being largely confined to the last 50 years. Given the paucity of long-term and direct measurements of the atmosphere and ocean, the description of long-term behavior of the climate system and the development of climate models for long-term simulation must rely on other sources of information. Put simply, we need other means to determine surface temperatures before there were thermometers. Fortunately there are other forms of data, known as *climate proxies*, that can have a statistical relationship with temperature and can been used to infer temperatures in the past before direct measurements. These proxies can give different representations of temperature at different temporal and spatial scales. Because of the heterogeneity among different kinds of proxies, most of the work in the paleoclimate community has focused on reconstructions that use a single type of climate proxy to infer temperature. Our goal is to extend reconstruction methods to exploit different kinds of climate proxies and thereby improve the accuracy of the reconstruction at different time scales.

This work proposes Bayesian hierarchical models (BHM) that are able to combine different kinds of climate proxies to create a single reconstruction for NH average temperature and also provide rigorous measures of uncertainty in the reconstructed temperatures. It is not obvious how to evaluate these methods for a temperature reconstruction where the true past temperatures are by no means available. In this paper we follow the convention in the paleoclimate community to evaluate reconstruction approaches using a high resolution simulation

Bo Li is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906 (E-mail: *boli@stat.purdue.edu*). Douglas W. Nychka is Senior Scientist and Director of Institute for Mathematics Applied to Geosciences (E-mail: *nychka@ucar.edu*) and Caspar M. Ammann is Scientist (E-mail: *ammann@ucar.edu*), National Center for Atmospheric Research (NCAR), Boulder, CO 80307. This research was supported by NCAR which is funded by the National Science Foundation. Additional support was provided through NSF CMG Collaborative Research award 0724828 and DMS-1007686. The authors thank the editor, the associate editor, and the referees for constructive suggestions that have improved the content and presentation of this article, also thank Marc Genton, Robert Harris, and Eugene Wahl for helpful discussions, and acknowledges the support through the NCAR Weather and Climate Impact Assessment Science Program and Linda Mearns.

© 2010 American Statistical Association
Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap09379

of the Earth's climate as "truth," and we also apply our methods to synthetic proxy data generated from this simulation. The synthetic proxy data bear the main characteristics of the real proxies.

A scientific contribution of this paper is an estimate of the added value of combining proxies with different climate retention characteristics and including external forcings series. To a statistician this is a practical exploration of different statistical designs for climate proxy and forcing datasets. Assembling a meaningful multiproxy dataset is by itself time consuming and thus an important question is whether such an effort is worth the additional accuracy in the reconstruction. A companion issue is the value of including external forcing information into the statistical procedure, especially when the forcing series also have observational errors. The inclusion of forcings can be criticized as adding nonclimatic information that should better be used in the post reconstruction to evaluate the cause of the climatic variations. However, if the forcings provide substantial improvement in the reconstruction and are included with relevant components of statistical uncertainty then they might be very useful. In particular, we are interested in determining the tradeoff between using external forcings jointly with a single proxy type or using multiple proxies. To our knowledge this is the first deliberate exploration of these statistical design/data issues for paleoclimate applications. Given the scope of these methodological issues we defer any reconstructions based on actual observations and their geophysical interpretation to a subsequent paper.

1.1 Multiple Proxies for Temperature

Different proxies preserve the climate information in different ways and therefore are sensitive to climate variables at different time scales. One might be good at short time scales, while another better at longer time scales. This is a key to our approach and the Bayesian hierarchical modeling that we employ takes advantage of such complementary skills/characteristics among different proxies. In this work we consider three widely used but distinct climate proxies: tree rings, borehole temperatures, and pollen abundance. Tree-ring width and density are perhaps the most widely distributed and generally used proxy and their relationship to seasonal temperatures has been extensively studied (Fritts 1976; Cook and Kairiukstis 1990; Schweingruber 1996). Although tree-ring measurements typically have their dating accurate to the year, their ability to encode centennial and slower climate variability is often limited by the technique used to remove nonclimatic variations in tree-ring time series (Cook et al. 1995; Briffa and Melvin 2008). Borehole depth temperature profiles directly preserve the surface temperature variability as the surface heat diffuses downward into the earth and have been recently used to characterize NH continental temperature for the past 500 years (Huang, Pollack, and Shen 2000; Harris and Chapman 2001; Chapman, Bartlett, and Harris 2004). As opposed to tree rings, borehole temperatures are only sensitive to climate variations at multidecadal or longer time scales due to the attenuation by the diffusion process. As a third proxy, pollen records provide climate information that can fill the gap between tree ring and borehole temperatures because they are considered sensitive to multidecadal variability. Although pollen records are widely distributed and have been used for reconstructing temperatures at discrete periods in the past (Guiot, Harrison, and Prentice 1993; Williams, Bartlein, and Webb 2000), they have only rarely been used in large-scale temperature reconstructions for the past millennium. For more details on these proxy characteristics, see Guiot et al. (2005).

1.2 External Drivers of Climate

Besides proxy measurements of past climate, there are also observations of the external drivers of the climate system. In this work we will focus on solar irradiance, volcanism, and greenhouse gases as primary external forcings for temperature evolution. These three forcings drive the large-scale climate variation, because the climate system has to react to any variations of these three forcings in order to restore the energy balance (Crowley 2000; IPCC 2007). A positive solar irradiance forcing tends to warm the surface, whereas a negative one tends to cool it. Volcanism often causes sudden temperature drops because the large amount of aerosols ejected by an explosive volcanic eruption into the atmosphere reduces the radiation reaching the surface. Unlike those two natural forcings, recent increase in greenhouse gases, with CO_2 as a major component, is a forcing due to human activities. An important component of our statistical model is the inclusion of an empirical model for temperature that depends on these forcings. We believe that this covariate information is an important addition to that from the proxy observations (Hughes and Ammann 2009), and also distinguishes our approach from many conventional paleoclimate reconstructions.

1.3 Statistical Estimates of Past Temperatures

Most studies reconstruct the past temperature by relying on only one proxy record of a particular resolution and do not use external forcings as covariate information. For example, Mann, Bradley, and Hughes (1998), Jones et al. (1998), and Crowley and Lowery (2000) reconstructed the NH temperature based on annually recorded proxy data, with Esper, Cook, and Schweingruber (2002) and Briffa and Melvin (2008) exclusively based on tree rings. Viau et al. (2006) carried out the reconstruction by focusing on pollen only and Harris and Chapman (2001) drew inference for the past temperature primarily based on borehole. Only a few reconstruction applications have tried to integrate data from sources with very different temporal resolution (e.g., Beltrami et al. 1995; Huang 2004; Moberg et al. 2005; Haslett et al. 2006). Although they found that data integration improves the reconstruction, none of them has integrated all three types of proxies as well as the external forcings mentioned above. Furthermore, no work has been conducted to systematically investigate the role of proxies and forcings and thus provide a guide for temperature reconstruction.

We develop a Bayesian hierarchical model (BHM) to reconstruct the NH mean temperature that incorporates the information from tree ring, pollen, and borehole records altogether, and also makes use of past forcings. BHMs have been demonstrated to be a powerful method in solving complex problems in climatology, ecology, and environmetrics (e.g., Wikle et al. 2001; Berliner, Milliff, and Wikle 2003) by splitting a complicated model into three basic components: an observation level, a process level, and a level of prior information on statistical parameters. In this application the observation level relates each

proxy to temperature, the process level relates temperature to the external forcings and the prior level specifies prior distributions of regression and variance parameters that tend to stabilize the problem but are otherwise uninformative. The synthesis of proxies and forcings enabled by BHMs will ideally provide more accurate reconstructions because the strength of one component can compensate for weakness of others.

Compared to the widely used regression approach (e.g., Li, Nychka, and Ammann 2007) in the temperature reconstruction, our method can avoid the possible attenuation effects caused by errors in explanatory variables (Ammann, Genton, and Li 2010), because the BHM explicitly models the measurement error in proxies. To our knowledge Haslett et al. (2006) and Lee, Zwiers, and Tsao (2008) are the two earliest statistical approaches that consider an observational model for the proxies. Haslett et al. first proposed a Bayesian reconstruction based on the fossil pollen data, but they did not consider other type of proxies and external forcings. Lee, Zwiers, and Tsao proposed a Kalman filter approach to incorporate the external driving factors, yet they did not consider combining proxies with different resolution.

1.4 Evaluation of Reconstruction Methods

In developing and testing this statistical method we take an unconventional approach in its evaluation of scientific impact. Because there is no adequate "ground truth" available for evaluating the fidelity of reconstruction approaches, we use a high-resolution, state-of-the-art climate simulation of the past 1150 years (Ammann et al. 2007) as a means for evaluating the method. Specifically we use the output from the climate model to generate synthetic proxy datasets that represent the characteristics of real-world proxies and include the forcings used in the simulation to determine how well our method can reconstruct the model temperatures. The strategy of using synthetic data from climate model output to evaluate the reconstruction methods is well established in the paleoclimatology literature (Mann and Rutherford 2002; Rutherford et al. 2003; Zorita, Gonzalez-Rouco, and Legutke 2003; von Storch et al. 2004; Ammann and Wahl 2007; Lee, Zwiers, and Tsao 2008) and is a practical solution to provide test beds that are complex but where the true temperatures are assumed known. Finally it should be noted that the climate model has substantial complexity relative to any tractable BHM and so studies using this model provide a reasonable measure of how well a statistical model can account for high dimensional and nonlinear geophysical process with stochastic components.

1.5 Outline of Article

Section 2 gives details on the global climate model output used for evaluating the method and the external forcings series for the past 1150 years. Section 3 describes the salient features of different proxy data and how we generate the synthetic proxies. Section 4 presents the BHM for combining proxies and forcings in temperature reconstruction, and discusses several variations of the hierarchical model. Section 5 shows the results from different hierarchical models under different combinations of proxies and forcings, answers the design questions raised in the Introduction, and analyzes the identifiability of parameters. Section 6 discusses the strength and extensibility of this Bayesian hierarchical framework with respect to temperature reconstruction. Finally, all the data, R files, and supplement material involved in this paper are posted at *http://www.image.ucar.edu/Data/*.

2. CLIMATE MODEL OUTPUT AND RADIATIVE FORCINGS

Climate system models are large computer codes that implement the basic physical equations for fluid dynamics and for thermodynamics to describe the motion of the atmosphere, ocean, and sea ice and their interaction with the land. The models are highly nonlinear and expressed in a differential form where the state of the climate system is stepped from one time point to the next over a short time interval by solving a large system of coupled partial differential equations. The model simulations are started by initial conditions of the ocean, atmosphere, and sea ice and are then subsequently driven by internal variability that is modulated or changed over time by external forcings such as solar irradiance, volcanic aerosols, and greenhouse gas concentrations. In addition, topography and an annual cycle of land cover are prescribed as boundary conditions for the atmosphere over land. Averaging the results of the model over a specific time period provides an estimate of the climate of the model.

The climate simulation used in this work (Ammann et al. 2007) is a run of the National Center for Atmospheric Research (NCAR) Community Climate System Model (CCSM) Version 1.4 (Boville et al. 2001; Otto-Bliesner and Brady 2001). The model atmosphere and land components were configured with a resolution of $3.75° \times 3.75°$ or 400 km $\times$ 400 km, a 3° resolution in the ocean with meridional resolution of $<1°$ at the equator. Details of this experiment can be found in Ammann et al. (2007) and at *http://www.cesm.ucar.edu*. Its simulated global annual temperature during 850–1999 is shown in Figure 1. A near hockey stick shape of this temperature series results from the increased forcings due to greenhouse gases. This shape, while a realistic representation of how temperature has increased, can cause a bias in the reconstruction and is discussed in the section on numerical results.

The three external forcings we consider here are solar irradiance, volcanism, and greenhouse gases. Those are also the key forcings used in the simulation by Ammann et al. (2007). For simplicity, we have not separated out tropospheric aerosol, as its variation after 1870 is similar to the greenhouse gases. The solar irradiance series is a reconstruction by Bard et al. (2000) and is derived from measurements of fluctuations of ^{10}Be production rates which is modulated by solar magnetic variability. The volcanic series is based on a synthesis of individual ice cores (see *http://www.ncdc.noaa.gov/paleo/icecore.html*) and in some cases on historical records of large eruptions. It is then transformed to be an estimate of volcanic sulfate aerosol mass (Ammann et al. 2007) whose amplitude reflects the radiation reduction. The change in greenhouse gases and its forcing prior to the 1950s is derived from air bubbles in ice cores and then direct measurements exist subsequently. We use carbon dioxide as a simple representation because other gases changed very similarly, albeit with different relative concentration. The shape of the carbon dioxide series is dominated by the first slow then rapid increase since the beginning of the 19th century. Figure 1 illustrates the three forcing series.

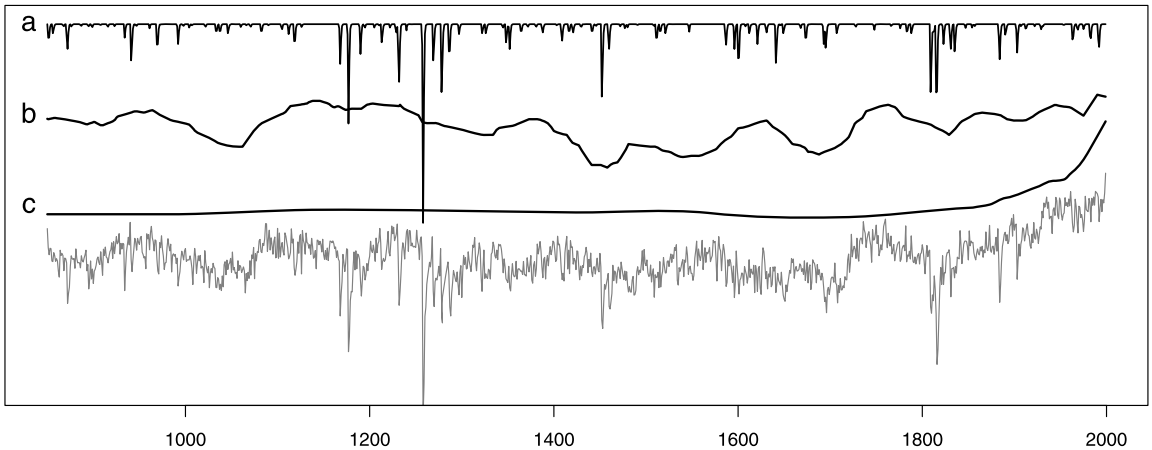

Figure 1. The diagram of three forcings. The grey curve is the NH temperature, and among the black curves (a) is the volcanism, (b) is the solar irradiance, and (c) is the greenhouse gases. All the curves are scaled in order to show them clearly in one figure.

Because the BHM will estimate the regression relationship between these variables and temperature, the individual scales of these series will not affect the reconstruction. This is useful as some of the uncertainty or error in these series is attributed to the absolute scale. For example, most discrepancies among different solar irradiance estimates are due to different scaling (Bard et al. 2000) which has little effect to our approach. The year-to-year variation in the greenhouse effect is very small for centuries before about 1800 and will have negligible effect because the key impact will more likely occur at the end of the time series. The estimates of volcanism, however, bears uncertainty in individual events that can reach 50% of the magnitude (Rind 1995 and Zielinski 2000), and thus needs to be taken into account by a statistical model.

3. CLIMATE PROXIES OBSERVATIONS AND GENERATING SYNTHETIC PROXIES

3.1 Tree Rings

Variation in the width and density of tree rings represent the most widely known climate proxy. Moreover, the wide geophysical distribution of trees make them well suited to high-resolution paleoclimate research. As a result of strong replication within and between specific sites and regions, and careful checking for common patterns, dating is effectively absolute, that is, accurate to the year. However, the potentially limited ability of capturing centennial and slower climate variability might inhibit the tree-ring networks from representing climate uniformly across the frequency spectrum, and hence the tree-ring-based reconstruction should be viewed with caution for time scales ranging from multidecadal to multicentennial (Cook et al. 1995).

Synthetic Tree-Ring Observations. To generate synthetic but realistic tree rings from the climate model simulation, we first select 15 local temperature series from the CCSM output and then generate 15 pseudo tree-ring records by a high pass filter. Specifically, we subtract the 10-year smoothing average from each of the local temperature series to give a filtered result. This might be an extreme approach for representing tree-ring information that perhaps in reality will not occur so drastically. However, the high frequency representation of tree rings in general has been well studied and accepted in the dendrochronology community (e.g., Cook et al. 1995 and Briffa et al. 1996), and the strategy of treating tree rings only being informative at high frequencies was originally employed by Moberg et al. (2005). So it serves the purpose of our study that evaluates the capability of our method in combining different proxies. Figure 2 shows the 15 locations and Figure 3 gives an example of generated tree rings.

3.2 Borehole Temperatures

Temperature-depth profiles measured in boreholes contain a record of surface temperature changes due to the thermal diffusion in the Earth, hence they provide a means to directly estimate the past temperatures through inversion of the down-core temperature profile after taking the natural geothermal gradient into account (e.g., Beltrami 2001). However, the borehole profile itself is unable to recover the surface temperatures at annual resolution, because the ground essentially behaves as a low pass filter only retaining the long term trends of climate. In particular, the filtering becomes more and more severe with time further back and thus smears out the potentially recoverable temperatures. As a result, borehole temperatures are only sensitive to climate variations at multidecadal or longer time scales (Gonzalez-Rouco, von Storch, and Zorita 2003; Mann and Schmidt 2003; Chapman, Bartlett, and Harris 2004; Huang 2004). Recently, borehole-derived temperature estimates have attracted much attention because several studies have systematically pointed towards a larger temperature change since AD 1500 over the NH continents compared to the other reconstructions (e.g., Huang, Pollack, and Shen 2000).

Synthetic Borehole Observations. Unlike generating synthetic tree-ring proxies using individual temperature series, we generate borehole data based on five regional composite temperature series. These five composite temperature series are the local average of model temperature output over five $20° \times 20°$ squares as shown in Figure 2. The distribution of those locations reasonably represents the spread of real borehole data (Huang, Pollack, and Shen 2000). Due to the complexity of the physical process in forming the borehole profile, the algorithm

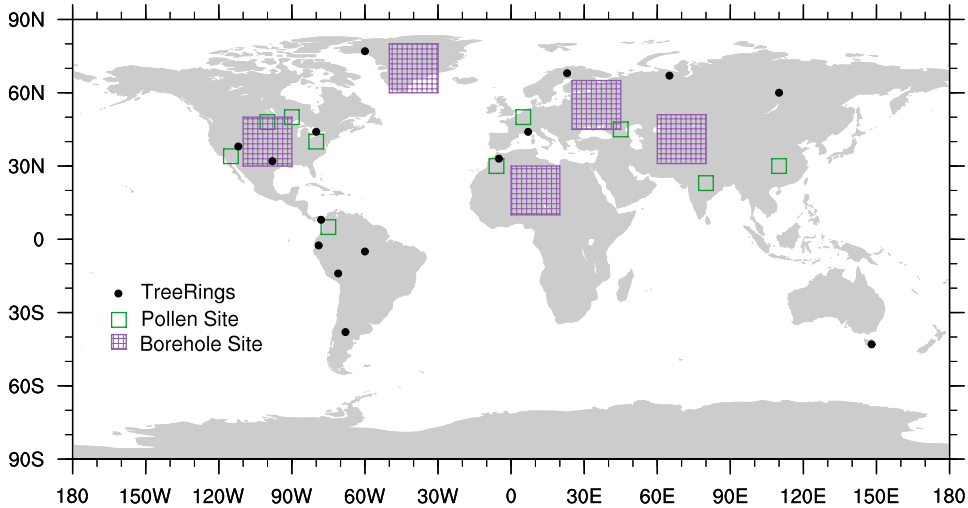

Figure 2. Pseudo-proxy sampling locations in the Community Climate System Model. The online version of this figure is in color.

to generate the borehole depth temperature is not straightforward. Therefore, we follow the modern preobservational mean-surface air temperature (POM-SAT) model which was originally derived by Carslaw and Jaegar (1959) and recently discussed by Harris and Chapman (2001) and Harris (2007) to simulate the borehole profiles up to 500 m. The value of the profile at every 5 m depth interval is considered the synthetic depth temperature. The POM-SAT model basically describes the diffusion process of surface temperature given an appropriate initial condition while having the attenuation of a thermal perturbation with respect to depth taken into account. This model has proven to be consistent with all diffusion processes (Harris 2007). To illustrate the characteristic of the borehole data, we show in Figure 4 artificial examples of how a 1000-year-long constant time series with a perturbation at different ages is represented in the borehole temperature profile. As expected for a diffusion process, the older a surface perturbation, the more smeared out in the borehole profile. For example, the temperature profile corresponding to the time series with the earliest pulse is the most flat among the five due to the severe smearing of this pulse, whereas the profile corresponding to the latest pulse displays a clear perturbation. As a consequence, the flat profile lost the most information about the exact timing, the nature and duration of the pulse.

3.3 Pollen Indices

Pollen assemblages retain a smoothed record of climate variation due to the persistence properties of mature plants (Brown et al. 2005), even where the data itself might be available at higher resolution. Since fossil pollen records possess skills in recovering temperatures at bidecadal to semicentennial time resolution, they provide climate information that can fill the gap between tree rings and borehole data (Bradley 1999).

Synthetic Pollen Observations. Similar to the procedure of generating synthetic borehole proxies, we select 10 regional composite temperature series as the local average of $7.5° \times 7.5°$ squares to generate pollen data. The locations of those squares are also shown in Figure 2. As discussed above that pollen carries information only at multidecadal temporal resolution, we

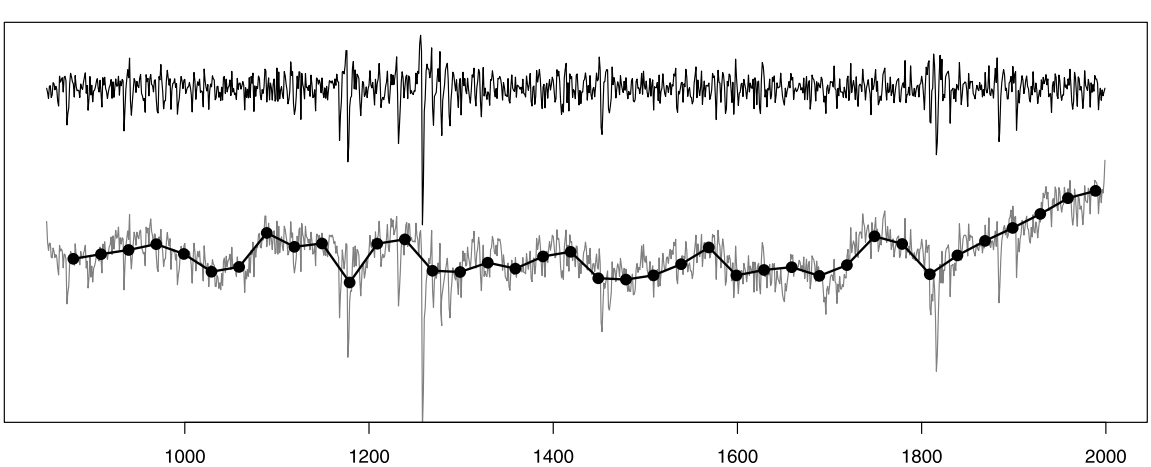

Figure 3. An example of synthetic tree rings and synthetic pollen together with the temperature (grey curve). The tree rings are represented by the upper black curve, and the pollen is shown by dots which are observed at every 30 years. The black curve with dots embedded is the 10-year smoothing average of the temperature.

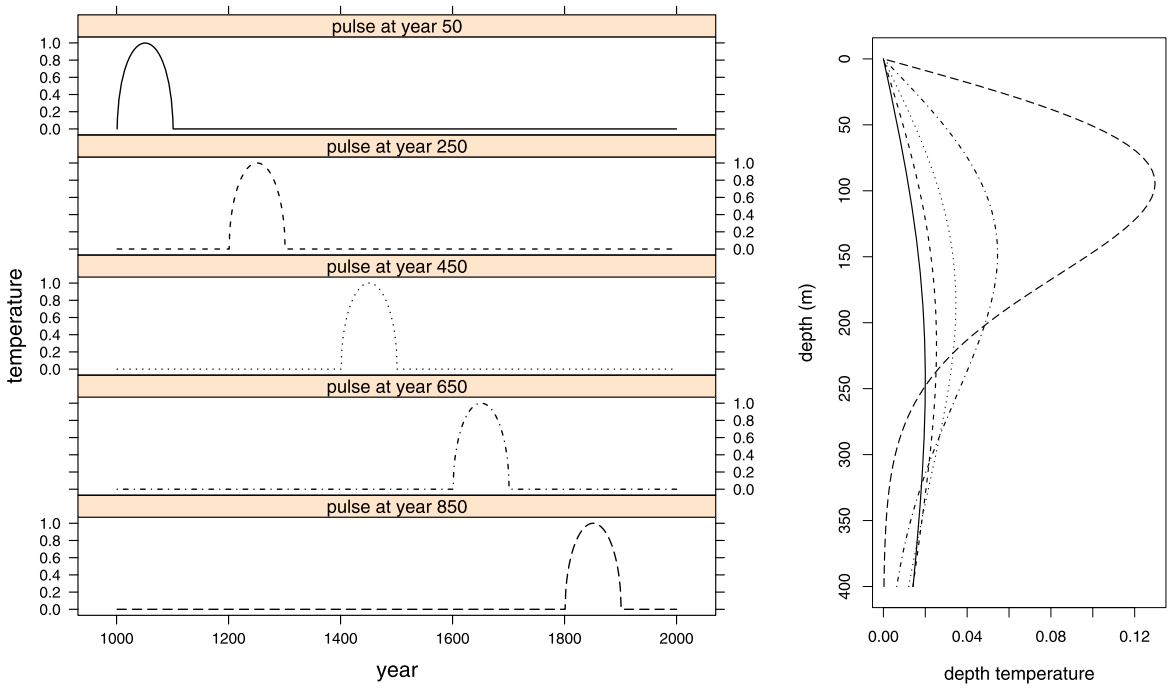

Figure 4. The borehole profile corresponding to the temperature series with a pulse at different time locations. This illustrates that the older a surface perturbation, the more smeared out in the borehole profile due to the diffusion. The online version of this figure is in color.

therefore mimic a pollen assemblage by sampling a 10-year average temperature series at 30-year intervals. The strategy of having pollen contain only lower frequency information than tree rings is analogous to the wavelet decomposition in Moberg et al. (2005). Figure 3 gives an example of such a generated pollen series.

Note that the above three temperature transformations in Sections 3.1 to 3.3 for generating proxies determine the three transformation matrices of $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$ in Section 4. More details of synthetic proxies generation can be seen in the supplement and the corresponding R files.

3.4 Noise in Proxies

Real world proxies are expected to contain extra noise in addition to the uncertainty between temperature at local scales and the hemispheric average. In order to assess the sensitivity of our approach to the noise in proxies, we additionally synthesize another set of proxies with an error component by adding white noise to the local/regional temperature series before they are processed to generate synthetic tree rings, pollen, and borehole. The variance of the noise is chosen to give a signal to noise ratio of 1 : 4 that conservatively reflects the expected precision in actual data (Mann et al. 2005). The reason that we add random perturbation to the original temperature rather than directly to the synthetic proxies is to preserve the smooth profile of pollen and borehole temperatures, and it more realistically represents the local climate noise that subsequently carries over into the proxy archives.

4. BAYESIAN HIERARCHICAL MODEL

BHMs split a complicated model into three basic components. The data model occupies one level of the hierarchy, while the process model resides below it. Typically, a third hierarchical level contains statistical models, also called priors, for unknown parameters that includes additional physical information. The levels are formally generated by a series of conditioning steps where one level is conditioned on knowledge of the levels below it. The reader is referred to Banerjee, Carlin, and Gelfand (2004) for an introduction to BHMs. Let $[X, Y]$ denote the joint probability density function of the random variables X and Y and $[X|Y]$ the conditional density of X *given* Y. Now let P denote proxy observations, T the NH temperature process and $\boldsymbol{\theta}$ a set of statistical parameters that are involved in specifying the joint distribution of P and T. The model specification is precisely the joint distribution $[P, T, \boldsymbol{\theta}]$. This form can be built from the product of conditional distributions:

$$[P, T, \boldsymbol{\theta}] = [P|T, \boldsymbol{\theta}][T|\boldsymbol{\theta}][\boldsymbol{\theta}].$$

Regarding the paleoclimate reconstruction problem an adumbration of the hierarchial levels is (i) Data stage $[P|T, \boldsymbol{\theta}]$, (ii) Process stage $[T|\boldsymbol{\theta}]$, (iii) Priors $[\boldsymbol{\theta}]$. Level (i) focuses on modeling statistical errors of the observed data and presents the likelihood of the observed proxies given the true temperature process, while level (ii) models the temperature process from the physical perspective. Level (iii) gives prior distributions of the unknown parameters and closes the hierarchy.

4.1 Full Model

Let $\mathbf{D}_i$, $\mathbf{P}_j$, and $\mathbf{B}_k$ be vectors of synthetic tree ring (**D**endrochronology), **P**ollen, and **B**orehole data indexed by their various locations. Note that these groups of proxy vectors will have different lengths due to their sampling. Moreover, each tree ring and pollen vectors are indexed with respect to time, and the borehole vectors are indexed by depth. Let $\mathbf{S}$, $\mathbf{V}_0$, and $\mathbf{C}$ be the

time series vectors of solar irradiance, volcanism, and greenhouse gases, and let $\mathbf{V}$ denote the volcanic series with error. Also let $\mathbf{1}$ denote a vector of ones with a generic length which will be determined individually by the conformable condition according to the local context.

Let $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$ be the three transformation matrices to link temperature series to tree ring, pollen, and borehole, respectively. Here we assume that those transformations are known. Although the real relationship between proxies and temperatures can be more complex than such a linear transformation, those three matrices are derived from the main characteristics of tree ring, pollen, and borehole data described in Section 3. More specifically, they represent the linear filters used to generate the corresponding pseudo proxies from the model temperature series. For example, the $\mathbf{M}_B$ is derived from the POMSAT model that is used to generate synthetic borehole profiles. Note that those matrices can be easily updated if more precise working models between proxies and temperatures are developed. Finally it is useful to partition the full-length temperature process $\mathbf{T}$ into the unknown temperatures $\mathbf{T}_1$ requiring reconstruction over the time span of available proxy data, and the observed instrumental temperatures $\mathbf{T}_2$ (1850–present), that is, $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')'$. Then we have the data and process models below:

(i) Data stage:

$$\mathbf{D}_i|(\mathbf{T}_1', \mathbf{T}_2')' = \mu_{iD}\mathbf{1} + \beta_{iD}\mathbf{M}_D(\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_{iD},$$
$$\boldsymbol{\epsilon}_{iD} \sim \mathrm{AR}(2)(\sigma_D^2, \phi_{1D}, \phi_{2D}), \tag{4.1}$$
$$\mathbf{P}_j|(\mathbf{T}_1', \mathbf{T}_2')' = \mu_{jP}\mathbf{1} + \beta_{jP}\mathbf{M}_P(\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_{jP},$$
$$\boldsymbol{\epsilon}_{jP} \sim \mathrm{AR}(2)(\sigma_P^2, \phi_{1P}, \phi_{2P}), \tag{4.2}$$
$$\mathbf{B}_k|(\mathbf{T}_1', \mathbf{T}_2')' = \mathbf{M}_B\{\mu_{kB}\mathbf{1} + \beta_{kB}(\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_{kB}\},$$
$$\boldsymbol{\epsilon}_{kB} \sim \text{iid } \mathrm{N}(0, \sigma_B^2), \tag{4.3}$$
$$\mathbf{V}|\mathbf{V}_0 = (1 + \epsilon_V)\mathbf{V}_0,$$
$$\epsilon_V \sim \text{iid } \mathrm{N}(0, 1/64). \tag{4.4}$$

(ii) Process stage:

$$(\mathbf{T}_1', \mathbf{T}_2')'|(\mathbf{S}, \mathbf{V}_0, \mathbf{C}) = \beta_0\mathbf{1} + \beta_1\mathbf{S} + \beta_2\mathbf{V}_0 + \beta_3\mathbf{C} + \boldsymbol{\epsilon}_T,$$
$$\boldsymbol{\epsilon}_T \sim \mathrm{AR}(2)(\sigma_T^2, \phi_{1T}, \phi_{2T}). \tag{4.5}$$

The target is to estimate $\mathbf{T}_1$ given $\mathbf{T}_2$, the proxies, and the forcings. Forward models (4.1) to (4.3) describe the statistical relationship between proxies and the true temperature process. Those models assume a stationary linear relationship between a local proxy and NH temperature over time and also conditional independency between proxies given true temperatures. The special form of model (4.3) respects the smooth feature of the borehole profile by applying the smooth filter $\mathbf{M}_B$ also to the error term $\boldsymbol{\epsilon}_{kB}$, because as opposed to tree rings and pollen assemblage the real borehole profile is a smooth curve. Proxies of the same type but from different locations, such as the 15 tree rings, 10 pollens, or 5 boreholes generated in Section 3, are allowed to have different regression coefficients including the intercepts and slopes, but they all share the same parameters in the error process to retain the parsimony of the whole model. This restriction is reasonable if the proxies are first standardized before being used, which is a standard method in paleoclimatology (Bradley and Jones 1993; Osborn and Briffa 2006).

Model (4.4) accounts for the uncertainty in the volcanism which is estimated to be between ±25% of the magnitude of the individual volcanic pulses themselves. Model (4.5) brings the physical understanding of temperature evolution based on energy balance theory into the reconstruction. Although our BHM contains only linear models, we could of course replace those linear models by more complicated ones. However, we found that the linear models suffice for our data, and the problem at hand, that is, reconstructing NH mean temperature (Mann et al. 2008). Our experience with related data suggests that an AR(1) structure is likely sufficient but we use an AR(2) model to provide additional flexibility to accommodate the possibility of more complex dependence in our application. An $\mathrm{AR}(2)(\sigma^2, \phi_1, \phi_2)$ process is defined as $e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \epsilon_t$, $\epsilon_t \sim \text{iid Normal}(0, \sigma^2)$. The choice of priors and the justification for the linear models with AR(2) error structure can be seen in the supplement.

4.2 Sampling From the Posterior Distribution

The specific choice of priors for time lag coefficients (ϕ_{1L}, ϕ_{2L}) with $L \in \{D, P, T\}$ guarantees their corresponding AR(2) process to be stationary and causal (Shumway and Stoffer 2006, chapter 3), and the conjugate priors for all the other parameters allow for an explicit full conditional posterior distribution for those parameters and $\mathbf{T}_1$. There is no closed form for the posterior distribution of time lag coefficients. Thus the posterior is sampled by alternating the Gibbs sampler, which is used for updating $\mathbf{T}_1$ and parameters with explicit full conditional distribution, and the Metropolis–Hasting (M–H) algorithm, which is used for updating autoregressive parameters. More specifically, we generate posteriors by Gibbs sampling for $\mathbf{T}_1$, $\mathbf{V}_0$, (μ_{iL}, β_{iL}) with $L \in \{D, P, B\}$, β_i with $i = 0, 1, 2, 3$ and σ_L^2 with $L \in \{D, P, B, T\}$, and generate posteriors by M–H for (ϕ_{1L}, ϕ_{2L}) with $L \in \{D, P, T\}$.

Whenever the M–H algorithm is used, the acceptance rate is tuned to be roughly between 25% and 50% to secure adequate mixing of posterior samples (Gelman, Roberts, and Gilks 1996). We choose the hyperparameters $\widetilde{\mu}_{iL} = 0$ and $\widetilde{\mu}_{iL}' = 1$ for $L \in \{D, P, B\}$, because this represents the ideal case when the local/regional temperatures are not biased against the NH temperature. For a similar reason, $\widetilde{\mu}_i$ is set to be 0, 1, 1, 1 for $i = 0, 1, 2, 3$. We found the results are robust to different choices of those hyperparameters. In order to let the data determine the final estimates of the regression coefficients, we choose a relatively wide variance $\widetilde{\sigma}_{iL}^2 = \widetilde{\sigma}_{iL}'^2 = 1$ for $L \in \{D, P, B\}$ and $\widetilde{\sigma}_i^2 = 1$ for $i = 0, 1, 2, 3$ to make the priors less informative. The hyperparameters $(\widetilde{q}_L, \widetilde{r}_L)$ with $L \in \{D, P, B, T\}$ are set to be (3, 1) which corresponds to relatively vague prior knowledge. The convergence check by starting sampling from different sets of initial values indicates that β_0, β_2, which is the scale parameter for zero-inflated volcanism, and σ_T^2 in the process model converge less well than the others. However, our main interest, the temperature reconstructions appear to be very insensitive to different initial values (see the supplement for details).

4.3 Simplifications of the Full BHM

In order to answer questions raised in the Introduction, we consider several simplifications of the full model. Here we list the different factors that will figure into our numerical study.

Temperature Process Model Without External Forcings. In order to identify the role of forcings in the reconstruction, we compare reconstructions with forcings and without forcings being incorporated. We form BHMs without forcings by omitting (4.4) and replacing model (4.5) in the process stage by

$$(\mathbf{T}_1', \mathbf{T}_2')' = \beta_0 \mathbf{1} + \boldsymbol{\epsilon}_T, \qquad \boldsymbol{\epsilon}_T \sim \mathrm{AR}(2)(\sigma_T^2, \phi_{1T}, \phi_{2T}).$$

This is equivalent to setting the regression coefficients equal to zero in the full process model. The discrepancy between results of this variation and the full model reveals the value of the external forcings at the process level.

An "Oracle" Proxy. Since we are dealing with climate model output that serves for our data samples, there is an opportunity to test the reconstruction skill of the BHM approach under optimal conditions, that is, if instead of local proxy data there would have existed long perfect instrumental series at the same locations. Therefore, we consider an *oracle* (after the legendary oracle in the temple at Delphi that provided divine answers to questions posed by visitors) proxy that contains full knowledge of the true temperatures at the gridbox locations. These idealized proxies can be interpreted as having thermometers back in time at these locations and the reconstruction error then is dominated by the skill of these locations in reproducing the hemispheric average. The oracle reconstruction is a benchmark because it is the best reconstruction from a selected sample of local/regional temperature series. To obtain the "oracle" reconstruction, we replace the three data models (4.1) to (4.3) by one single model

$$\mathbf{T}_l | (\mathbf{T}_1', \mathbf{T}_2')' = \mu_l \mathbf{1} + \beta_l (\mathbf{T}_1', \mathbf{T}_2')' + \boldsymbol{\epsilon}_l, \qquad \boldsymbol{\epsilon}_l \sim \mathrm{AR}(2)(\sigma^2, \phi_1, \phi_2),$$

where $\mathbf{T}_l$ are the grid box temperature time series (30 in total) that are used to generate all three types of proxies. The other levels in the hierarchy are kept the same.

Subsets of Proxies. Although the oracle experiment provides a baseline reference, our main interest is in separating the contribution of each proxy to the reconstruction, and examining how deficiencies in one could be compensated for by others. This can be accomplished by reducing the full model into a sequence of submodels that only contain a subset of the three proxies, that is, this can be carried out by omitting different sets of proxies from the data level.

Partial Temperature Process Model. The temperature reconstruction usually assumes a stationary relationship between temperatures and proxies across time, and this is also implied in our hierarchical models. In addition, we assume that the temperature process at the prediction period $\mathbf{T}_1$ and at the calibration period $\mathbf{T}_2$ follow the same model. Thus it imposes the constraint that the mean function of temperatures, either as a function of forcings or as a constant, is identical for both the calibration and prediction periods. One can relax this constraints by modeling only $\mathbf{T}_1$ in the process stage, and investigate the effects caused by this assumption. In this case, the process model (4.5) becomes

$$\mathbf{T}_1 | (\mathbf{S}, \mathbf{V}_0, \mathbf{C}) = \beta_0 \mathbf{1} + \beta_1 \mathbf{S} + \beta_2 \mathbf{V}_0 + \beta_3 \mathbf{C} + \boldsymbol{\epsilon}_T, \qquad \boldsymbol{\epsilon}_T \sim \mathrm{AR}(2)(\sigma_T^2, \phi_{1T}, \phi_{2T}).$$

5. NUMERICAL STUDY

Motivated by the issues raised in the Introduction we evaluated the following five different subsets of proxy or related data.

(T): Oracle proxies,
(D): Tree ring (Dendrochronology) only,
(DP): Tree ring + Pollen,
(DB): Tree ring + Borehole,
(DBP): Tree ring + Borehole + Pollen.

For each of these cases we explored the different BHM choices with a 2^3 factorial design: with/without forcing covariates, with/without proxy noise and modeling $\mathbf{T}_1/\mathbf{T}$ in the process model. Thus our study consists of a total of 5×2^3 separate reconstructions that can be compared to the actual model temperatures. The bias and the variance of differences between the reconstruction and target are two informative measures to evaluate the reconstruction with the square root of the sum of the squared bias and the variance being the root mean squared error (rmse). Both low bias and low variance are desirable for being a good reconstruction, and we report the bias, variance, and rmse for all reconstructions in Figure 5. In general the patterns are what one can expect with the rmse decreasing as proxies are added and the oracle proxy having the smallest rmse.

In order to more formally compare different reconstructions, we employ the posterior predictive loss criterion proposed in Gelfand and Ghosh (1998) that considers predictive biases penalized by predictive variances. Specifically, we consider a loss function $D_k(m)$ for each model (reconstruction) m, $m = 1, 2, \ldots, 40$,

$$D_k(m) = P(m) + \frac{k}{k+1} G(m),$$

where $k \geq 0$ is a weighting parameter, $P(m)$ is the sum of predictive variances that imposes penalty on the complexity of models and $G(m)$ is the sum of squared errors that measures the goodness of fit. Letting $k \to \infty$, we have $D(m) = P(m) + G(m)$. For each of the 40 models, we studied $D_k(m)$ for $k = 1, 3, 9$, and ∞. We found that the ordering of models under $D_k(m)$ is insensitive to the choice of k, and moreover, the pattern of $D(m)$ highly matches the pattern of rmse in Figure 5(c) (see supplement). Therefore, for convenience we simply examine rmse for model comparison.

5.1 The Value of Forcings and Proxies

Influence of External Forcings. Forcings play a very important role in obtaining a well calibrated and sharp reconstruction. As seen in Figure 5, in all cases inclusion of forcing covariates reduce the bias, the variance and thus rmse. In particular, when there is no pollen data involved, forcings dramatically improve the performance of the reconstruction based on tree rings or a combination of tree rings and borehole. Yet once pollen is being employed, the benefit of incorporating forcings becomes less remarkable. Again, this is geophysically not surprising because pollen data contain the variability information of decadal to centennial scale. In short, external forcings are helpful if the included proxy data do not well represent the low frequency variability.

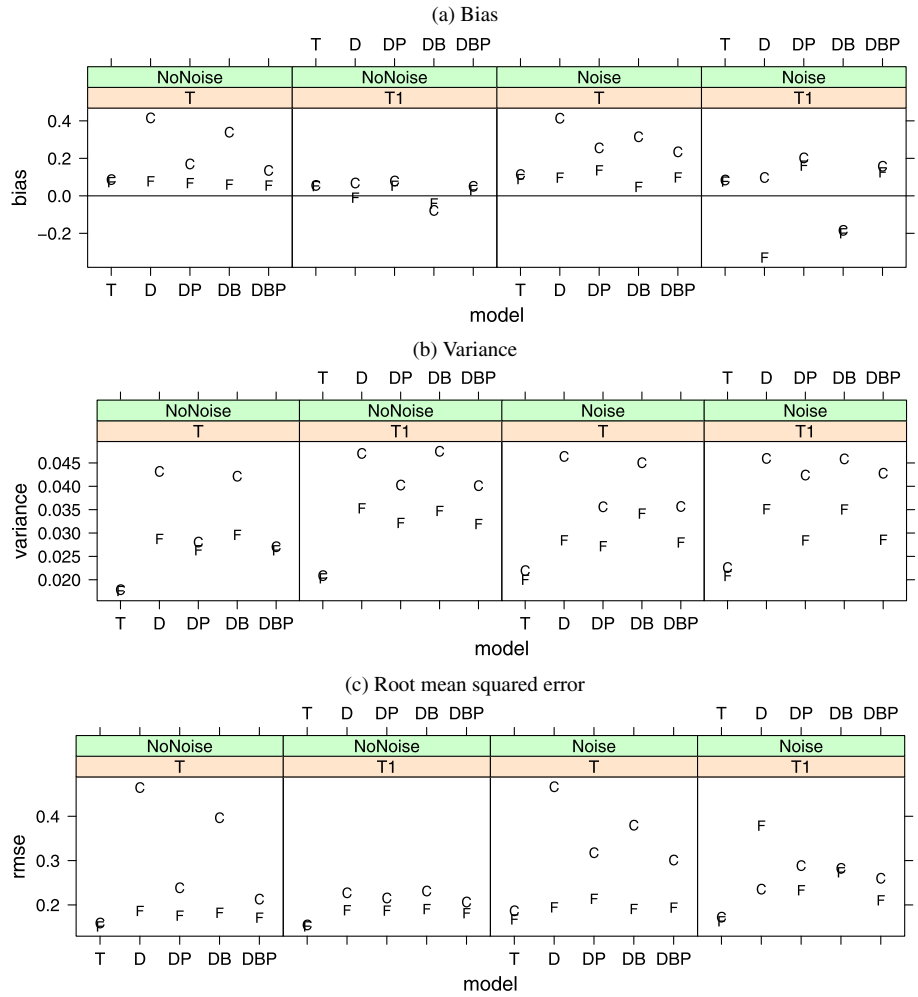

Figure 5. Bias, variance, and rmse of the reconstructions for five data models and 2^3 scenarios that are combinations of with/without forcings, with/without noise and modeling $\mathbf{T}_1/\mathbf{T}$ in the process stage. "C" and "F" are the reconstructions without forcings (with constant mean function) and with forcings incorporated, respectively. The online version of this figure is in color.

Skill of Proxies. We focus on the five data models (T, D, DP, DB, DBP) where forcings are absent, proxies are not subject to noise, and constant mean process model is assumed for $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')'$ rather than only $\mathbf{T}_1$, to study the contribution due to each type of proxies. This corresponds to the "C" points in the leftmost panels of Figures 5(a) to (c), and the patterns in those plots roughly imply the role of each type of proxy. To more clearly illustrate the benefit of incorporating a particular type of proxy, we compare the spectrum of the reconstruction residuals from the five different data subsets. The residual spectrum at different frequencies measures the variation component that we missed in the reconstruction at that specific frequency. Figure 6 shows that the spectrum based on data models (D) and (DB) look similar and also the spectrum based on data models (DP) and (DBP) are hardly distinguishable, although the latter two have smaller power at low frequencies due to the involvement of pollen proxies in the reconstruction. Since pollen was sampled every 30 years from a 10-year smoothing average of temperatures, it would thus be expected to retain the variability at around a 30-year period. This is verified by the observation that the spectrum of (DP) and (DBP) departures from the spectrum of (D) and (DB) at about a 30-year period after an agreement at high frequencies. Borehole information does not appear to contribute much because the spectrum only describes the variation whereas the information therein is smeared too much to recover any detail of temperature evolution other than a long term trend. However, as seen in Figure 5(a) the borehole does help to reduce the bias a bit. Overall, compared to the ora-

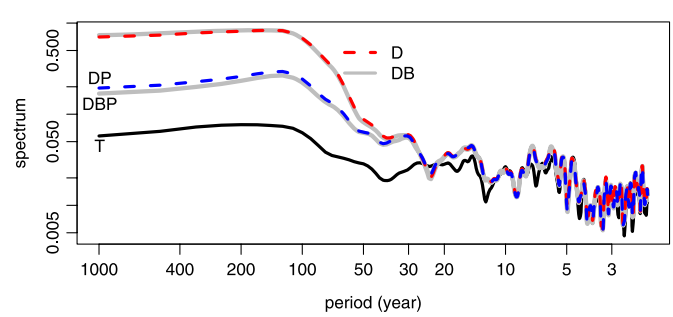

Figure 6. Using smoothed spectrum of reconstruction residuals from the five data models to illustrate the frequency band at which proxies capture the variation of the temperature process. Both axes are plotted on a logarithmic scale. The online version of this figure is in color.

cle proxies (T), the reconstructions using other proxies perform well at high frequency but are less precise at low frequency, as part of the low frequency information is lost in pollen and borehole.

5.2 Other Inferences

Cause of Bias. It can be seen from Figure 5(a) that the reconstruction when $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')'$ is modeled in the process stage carries some systematic positive bias. This positive bias is pronounced in cases where forcings or information at longer time scales is not available, that is, $\mathbf{T}$ is modeled as an AR(2) with constant mean which corresponds to the "C" points in the figure, and no pollen is used in the reconstruction. The reason for the positive bias is because the observed $\mathbf{T}_2$ which serves the primary source to estimate the mean of $\mathbf{T}$ has higher mean temperature than $\mathbf{T}_1$, but was nonetheless assumed to have the same mean function as $\mathbf{T}_1$ in the model. Therefore, after we remove this a priori assumption and only leave $\mathbf{T}_1$ in the process stage as described in the partial temperature process model in Section 4.3, the positive bias has been largely reduced. Another way to reduce the bias is to incorporate external forcings since this enables the temperature process to be estimated by its dependency on forcings. In this way the difference in mean functions is accounted for by the varying external forcings.

If bias is the primary concern, one should consider the reconstruction that is based on modeling only $\mathbf{T}_1$ in the process stage, though one would have to accept that it will carry more variance. Note that the oracle reconstruction is not necessarily the optimal reconstruction just in terms of bias. We can see that some reconstructions have even lower bias than their corresponding oracle experiment. Yet due to the bias–variance trade off, the oracle reconstruction is the best in terms of rmse. This is particularly visible when realistic noise is applied into proxies.

Sensitivity to Noise in Proxies. Not surprisingly, the noise in proxies introduces both bias and more variability in the reconstructions. We can see from Figure 5 that with the substantial amount of noise added to the proxies, the performance of the reconstruction deteriorates some but not terribly. It is worth noting that our results are only based on one set of contaminated proxies. Given different noise, the performance is slightly different, but does not appreciably change the basic conclusion.

5.3 Posterior Samples of the Reconstructed Temperature

We select model (DP) with forcings to show examples of the reconstruction at three scenarios. Figure 7(a) is for modeling $\mathbf{T}$ and no noise, Figure 7(b) for modeling $\mathbf{T}_1$ and no noise, and Figure 7(c) for modeling $\mathbf{T}$ and with noise. To make this comparison clearer we report posterior reconstructions for decadal average temperature. In all those figures, the 95% uncertainty band calculated from the posterior samples is also displayed together with the reconstruction. In Figure 7, all the reconstructions follow the trend of the target very well, although they seem to miss some details. Comparison between panels (a) and (b) in this figure shows that by modeling only the unknown $\mathbf{T}_1$ in the process stage can effectively reduce the bias caused by assuming $\mathbf{T}_1$ and $\mathbf{T}_2$ to have the same mean function. The difference between panels (a) and (b) and panel (c) illustrates the larger bias and wider uncertainty band introduced by noise. In panels (a) and (b), the uncertainty band covers the target temperatures fairly well, while in panel (c) the coverage deteriorates due to the bias.

In addition, we formally assess the model adequacy with posterior samples. Using the criterion of verification rank histogram and coverage probability of posterior distributions as in Gel, Raftery, and Gneiting (2004), we assess our models by comparing the performance of the reconstructions that are based on synthetic proxies to the reference reconstruction that is

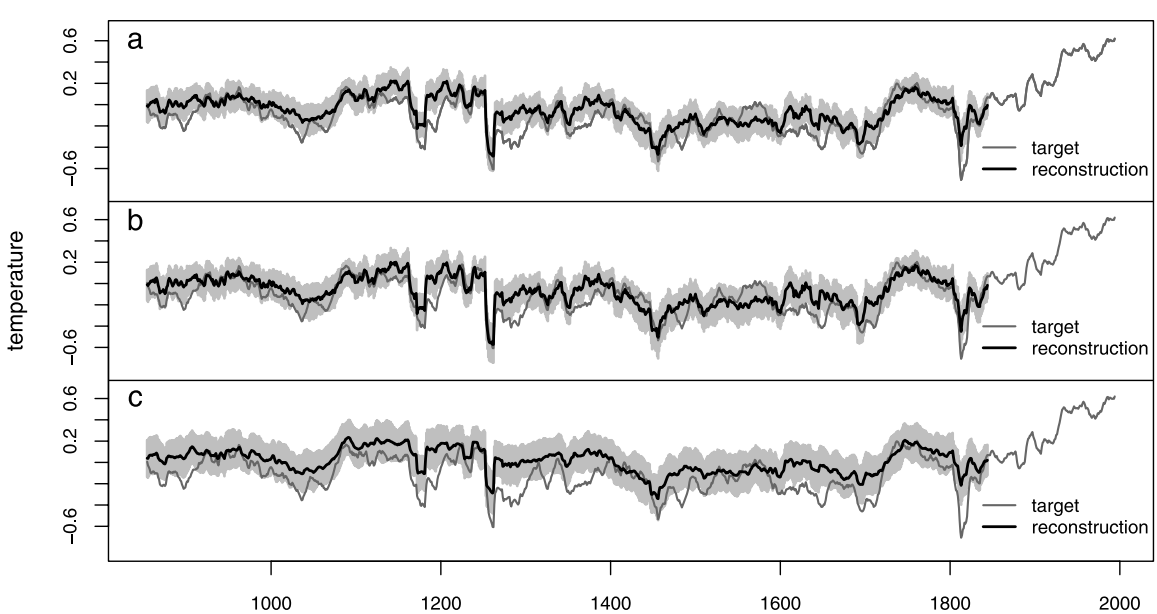

Figure 7. The reconstructions using tree rings and pollen together with forcings in three scenarios. (a) modeling $\mathbf{T}$ and without noise; (b) modeling $\mathbf{T}_1$ and without noise; (c) modeling $\mathbf{T}$ and with noise. The grey area is the 95% uncertainty band of the reconstruction.

based on "oracle" proxies. This is because the reference reconstruction reaches the capacity of those "oracle" proxies (local temperatures) in recovering the NH temperature and thus can serve as a baseline to evaluate the models for synthetic proxies. The results show no evidence towards inadequacy of those models (see details in the supplement).

5.4 Identifiability of Parameters

We examine the posterior distribution of parameter estimates and compare them to their corresponding priors to make sure that priors only have little influence on the parameter estimates. We focus on the model (DBP) with forcings included and with proxy errors, because this case has the most complex setting and provides the greatest challenge in determining parameters. We found that in general, the posteriors are not sensitive to the priors and this is suggested by the relatively small variance of the posteriors compared to their priors. However, compared to the stable estimates for regression coefficients for tree rings, the estimates for borehole data contain more uncertainties. Moreover, the BHM is unable to resolve the variance parameter in the borehole model. We conjecture that this is because the transformation matrix $\mathbf{M}_B$ is an ill-posed matrix that has a small effective rank, on the order of 6 degrees of freedom, hence there is no way to fully recover the temperature information from the borehole profile. This essentially has been illustrated in Figure 4. Despite the difficulty in estimating some of the parameters related to the borehole proxy data the resulting reconstructions are about the same as (DP) combinations or slightly better.

6. DISCUSSION AND CONCLUSIONS

This paper has proposed a new application of BHMs to reconstruct NH temperatures by jointly using proxy data with different temporal resolutions, and using forcings as external covariates of temperature evolution. With this method we investigated the benefits by combining different proxies and by inclusion of the forcings. Our results showed that a process model that includes the external forcings in the form of an energy balance can dramatically improve the reconstruction, particularly if the applied proxy data is deficient of decadal or centennial scale variability. In our numerical study this improvement can be by a factor of 2 in rmse. However, its role can be partially replaced by our hypothetical pollen proxy that fills this range. Tree rings play a significant role in retaining the high frequency variability, while pollen improved the reconstruction remarkably by capturing the variation at lower frequency band. These results make a case for attempting multiproxy reconstructions with tree rings and pollen assemblages and also including external forcing covariates. Although we base these conclusions on a synthetic Monte Carlo experiment, the climate model simulation used as truth is a complex and extensive representation of the actual climate system and so provides confidence that these results will extend well to real world conditions.

Pollen proxies are usually collected from sediments layers and thus possibly subject to dating errors, that is, the date during which pollen was formed or the age of a layer in lake sediment might not be exactly identified due to various reasons, such as the time lag between initial plant introduction to its abundance and different sediment accumulation rates (see Bradley 1999). We account for the dating error to some extent by considering additive proxy errors. However, one area of future work is to consider a more realistic model for dating errors (Haslett and Parnell 2008). One surprise in this work is that the hypothetical borehole proxies do not improve the reconstruction in a significant way. Nevertheless, in reality pollen records might not perform as well as the synthetic ones, and in such a case borehole information may be used more effectively. Also the use of borehole information might be improved by greater attention to the ill-posed aspects of the data model.

We believe that our work is a positive contribution to the paleoclimate community as it attempts to exploit all available proxy information to discern past climate. Also, whenever a Bayesian approach is used it will include companion measures of the reconstruction uncertainty. One perspective of the debate concerning the Northern Hemisphere temperature reconstruction proposed by Mann, Bradley, and Hughes (1998) is that most attention is centered on the estimate but only little on its uncertainty (see NRC 2006 for a comprehensive report). The "hockey stick" shape, which was the center of that discussion, in our context is an approximation to the posterior mean and then any assessment of the shape must also include the uncertainty about this estimate. The concept of an *ensemble* of possible states is a well-recognized technique in geosciences to quantify uncertainty in an estimate. Via Bayesian analysis, a random sample of states from the posterior provides a rigorous and easily interpreted method for generating an ensemble.

Another advantage of the BHM framework is that it readily extends to more complex data level or process models. Because of this flexibility we believe that the methods will adapt to more complex statistical features of real data and the experience in this study will transfer to more complicated cases. An important extension will be reconstructing the space–time temperature process instead of the NH temperature based on the work of Tingley and Huybers (2010), and further reconstructing the multivariate space–time climate process. Climate variables of interest include temperature, precipitation and geopotential heights. The Bayesian framework in this article naturally lends itself to univariate and multivariate space–time random field reconstruction. Achieving this long-term goal will provide a valuable analysis to evaluate the next generation of climate system models and improve our understanding of past climate.

SUPPLEMENTAL MATERIALS

Additional analysis details and results: More details and results of the analysis are shown in pdf file. (supplement.pdf)

Data and R codes: The tar file contains all data sets that have been used in reconstructions and all R codes that implement the reconstructions. The readme.rtf enclosed in the tar file describes the content of each data file and R code file. (paleo.tar)

[Received June 2009. Revised March 2010.]

REFERENCES

Ammann, C. M., and Wahl, E. (2007), "The Importance of the Geophysical Context in Statistical Evaluations of Climate Reconstruction Procedures," *Climatic Change*, 85, 71–88. [885]

Ammann, C. M., Genton, M. G., and Li, B. (2010), "Correcting for Signal Attenuation From Noisy Proxy Data in Climate Reconstruction," *Climate of the Past*, 6, 273–279. [885]

Ammann, C. M., Joos, F., Schimel, D., Otto-Bliesner, B. L., and Tomas, R. (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. [885]

Banerjee, S., Carlin, B., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton: Chapman & Hall/CRC. [888]

Bard, E., Raisbeck, G., Yiou, F., and Jouzel, J. (2000), "Solar Irradiance During the Last 1200 Years Based on Cosmogenic Nuclides," *Tellus, Ser. B*, 52, 985–992. [885,886]

Beltrami, H. (2001), "Surface Heat Flux Histories From Inversion of Geothermal Data: Energy Balance at the Earth's Surface," *Journal of Geophysical Research*, 106, 21979–21993. [886]

Beltrami, H., Chapman, D. S., Archambault, S., and Bergeron, Y. (1995), "Reconstruction of High Resolution Ground Temperature Histories Combining Dendrochronological and Geothermal Data," *Earth Planetary Science Letters*, 136, 437–445. [884]

Berliner, L. M., Milliff, R. F., and Wikle, C. K. (2003), "Bayesian Hierarchical Modeling of Air–Sea Interaction," *Journal of Geophysical Research*, 108, DOI: *10.10292002JC001413*. [884]

Boville, B. A., Kiehl, J. T., Rasch, P. J., and Bryan, F. O. (2001), "Improvements to the NCAR CSM-1 for Transient Climate Simulations," *Journal of Climate*, 14, 164–179. [885]

Bradley, R. S. (1999), *Paleoclimatology: Reconstructing Climates of the Quaternary* (2nd ed.), San Diego: Academic Press. [887,893]

Bradley, R. S., and Jones, P. D. (1993), ""Little Ice Age" Summer Temperature Variations: Their Nature and Relevance to Recent Global Warming Trends," *The Holocene*, 3, 367–376. [889]

Briffa, K. R., and Melvin, T. M. (2008), "A Closer Look at Regional Chronology Standardisation of Tree-Ring Records: Justification of the Need, a Warning of Some Pitfalls, and Suggested Improvements in Its Application," in *Dendroclimatology: Progress and Prospects. Developments in Paleoenvironmental Research*, eds. M. K. Hughes, H. F. Diaz, and T. W. Swetnam, New York: Springer. [884]

Briffa, K. R., Jones, P. D., Schweigruber, F. H., Karlén, W., and Shiyatov, S. G. (1996), "Tree-Ring Variables as Proxy-Climate Indicators: Problems With Low-Frequency Signals," in *Climatic Variations and Forcing Mechanism of the Last 2000 Years. Series I: Global Environmental Change*, eds. P. D. Jones, R. S. Bradley, and J. Jouzel, Berlin, Heidelberg: Springer-Verlag, pp. 9–41. [886]

Brown, K. J., Clark, J. S., Grimm, E. C., Donovan, J. J., Mueller, P. G., Hansen, B. C. S., and Stefanova, I. (2005), "Fire Cycles in North American Interior Grasslands and Their Relationship to Prairie Drought," *Proceedings of the National Academy of Sciences*, 102 (25), 8865–8870. [887]

Carslaw, H. S., and Jaeger, J. C. (1959), *Conduction of Heat in Solids* (2nd ed.), Oxford, U.K.: Oxford University Press. [887]

Chapman, D. S., Bartlett, M. G., and Harris, R. N. (2004), Comment on "Ground vs. Surface Air Temperature Trends: Implications for Borehole Surface Temperature Reconstructions" by M. E. Mann and G. Schmidt, *Geophysical Research Letters*, 31, L07205, DOI: *10.1029/2003GL019054*. [884,886]

Cook, E. R., and Kairiukstis, L. A. (1990), *Methods of Dendrochronology*, Dordrecht: Kluwer Academic. [884]

Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A., and Funkhouser, G. (1995), "The "Segment Length Curse" in Long Tree-Ring Chronology Development for Paleoclimatic Studies," *The Holocene*, 5, 229–237. [884,886]

Crowley, T. J. (2000), "Causes of Climate Change Over the Last 1000 Years," *Science*, 289, 270–277. [884]

Crowley, T. J., and Lowery, T. S. (2000), "How Warm Was the Medieval Warm Period?" *Ambio*, 29, 51–54. [884]

Esper, J., Cook, E. R., and Schweingruber, F. H. (2002), "Low-Frequency Signals in Long Tree-Ring Chronologies for Reconstructing Past Temperature Variability," *Science*, 295, 2250–2253. [884]

Fritts, H. C. (1976), *Tree Rings and Climate*, New York: Academic Press. [884]

Gel, Y., Raftery, A. E., and Gneiting, T. (2004), "Calibrated Probabilitic Mesoscale Weather Field Forecasting: The Geostatistical Output Perturbation Method," *Journal of the American Statistical Association*, 99, 575–583. [892]

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11. [890]

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996), "Efficient Metropolis Jumping Rules," *Baysian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press. [889]

Gonzalez-Rouco, F., von Storch, H., and Zorita, E. (2003), "Deep Soil Temperature as Proxy for Surface Air-Temperature in a Coupled Model Simulation of the Last Thousand Years," *Geophysical Research Letters*, 30 (21), 2116, DOI: *10.1029/2003GL018264*. [886]

Guiot, J., Harrison, S. P., and Prentice, C. I. (1993), "Reconstruction of Holocene Precipitation Patterns in Europe Using Pollen and Lake Level Data," *Quaternary Research*, 40, 139–149. [884]

Guiot, J., Nicault, A., Rathgeber, C., Edouard, J. L., Guibal, F., Pichard, G., and Till, C. (2005), "Last-Millennium Summer-Temperature Variations in Western Europe Based on Proxy Data," *The Holocene*, 15 (4), 489–500. [884]

Harris, R. N. (2007), "Variations in Air and Ground Temperature and the POMSAT Model: Results From the Northern Hemisphere," *Climate of the Past*, 3, 611–621. [887]

Harris, R. N., and Chapman, D. S. (2001), "Mid-Latitude (30–60N) Climatic Warming Inferred by Combining Borehole Temperatures With Surface Air Temperatures," *Geophysical Research Letters*, 28, 747–750. [884,887]

Haslett, J., and Parnell, A. (2008), "A Simple Monotone Process With Application to Radiocarbon-Dated Depth Chronologies," *Journal of the Royal Statistical Society, Ser. C*, 57, 399–418. [893]

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, Ser. A*, 169, 395–438. [884,885]

Huang, S. (2004), "Merging Information From Different Resources for New Insights Into Climate Change in the Past and Future," *Geophysical Research Letters*, 31, L13205. [884,886]

Huang, S., Pollack, H. N., and Shen, P. Y. (2000), "Temperature Trends Over the Past Five Centuries Reconstructed From Borehole Temperatures," *Nature*, 403, 756–758. [884,886]

Hughes, M. K., and Ammann, C. M. (2009), "The Future of the Past: An Earth System Framework for High Resolution Paleoclimatology: Editorial Essay," *Climate Change*, 94, 247–259. [884]

IPCC (2007), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, eds. S. Solomon, D. Quin, M. Manning, Z. Chen, M. Marquis, K. B. M. Tignor, and H. L. Miller, Cambridge, U.K. and New York: Cambridge University Press. [883,884]

Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B. (1998), "High-Resolution Palaeoclimatic Records for the Last Millennium: Interpretation, Integration and Comparison With General Circulation Model Control-Run Temperatures," *The Holocene*, 8, 455–471. [884]

Lee, T. C. K., Zwiers, F. W., and Tsao, M. (2008), "Evaluation of Proxy-Based Millennial Reconstruction Methods," *Climate Dynamics*, 31 (2–3), 263, DOI: *10.1007/s00382-007-0351-9*. [885]

Li, B., Nychka, W. D., and Ammann, C. M. (2007), "The "Hockey Stick" and the 1990s: A Statistical Perspective on Reconstructing Hemispheric Temperatures," *Tellus*, 59, 591–598. [885]

Mann, M. E., and Rutherford, S. (2002), "Climate Reconstruction Using "Pseudoproxies"," *Geophysical Research Letters*, 29, 1501, DOI: *10.1029/2001GL014554*. [885]

Mann, M. E., and Schmidt, G. A. (2003), "Ground vs. Surface Air Temperature Trends: Implications for Borehole Surface Temperature Reconstructions," *Geophysical Research Letters*, 30 (12), 1607, DOI: *10.1029/2003GL017170*. [886]

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. [884,893,895]

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C. (2005), "Testing the Fidelity of Methods Used in Proxy-Based Reconstruction of Past Climate," *Journal of Climate*, 18, 4097–4107. [888]

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., and Rutherford, S. (2008), "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. [889]

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. [884,886,888]

National Research Council (NRC) (2006), *Surface Temperature Reconstructions for the Last 2,000 Years*, Washington, DC: National Academy Press. [893]

Osborn, T. J., and Briffa, K. R. (2006), "The Spatial Extent of 20th-Century Warmth in the Context of the Past 1200 Years," *Science*, 311, 841–844. [889]

Otto-Bliesner, B. L., and Brady, E. C. (2001), "Tropical Pacific Variability in the NCAR Climate System Model," *Journal of Climate*, 14, 3587–3607. [885]

Rind, D. (1995), "The Potential for Modeling the Effects of Different Forcing Factors on Climate During the Past 2000 Years," in *Global Environmental Change. NATO ASI Series*, Vol. 41, ed. P. D. Jones, Stuttgart: Springer-Verlag, pp. 563–581. [886]

Rutherford, S., Mann, M. E., Delworth, T. L., and Stouffer, R. (2003), "Climate Field Reconstruction Under Stationary and Nonstationary Forcing," *Journal of Climate*, 16, 462–479. [885]

Schweingruber, F. H. (1996), *Tree Rings and Environment Dendroecology*, Berne, Switzerland: Paul Haupt. [884]

Shumway, R. H., and Stoffer, D. S. (2006), *Time Series Analysis and Its Applications: With R Examples*, New York: Springer. [889]

Tingley, M. P., and Huybers, P. (2010), "A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and Applications to Paleoclimate Reconstruction Problems," *Journal of Climate*, 23, 2759–2781. [893]

Viau, A. E., Gajewski, K., Sawada, M. C., and Fines, P. (2006), "Millennial-Scale Temperature Variations in North America During the Holocene," *Journal of Geophysical Research*, 111, D09012, DOI: *10.1029/2005JD006031*. [884]

von Storch, H., Zorita, E., Jones, J., Dimitriev, Y., Gonzalez-Rouco, F., and Tett, S. (2004), "Reconstructing Past Climate From Noisy Data," *Science*, 306, 679–682. [885]

Wikle, C. K., Millif, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatiotemporal Hierarchical Bayesian Modeling: Tropical Ocean Surface Winds," *Journal of the American Statistical Association*, 96, 382–397. [884]

Williams, J. W., Bartlein, P. J., and Webb III, T. (2000), "Data-Model Comparisons for Eastern North America—Inferred Biomes and Climate Values From Pollen Data," in *Proceedings of the Third Paleoclimatic Modeling Intercomparison Project Workshop*, ed. P. Braconnot, Montreal, Canada: World Climate Research Program, pp. 77–86. [884]

Zielinski, G. A. (2000), "Use of Paleo-Records in Determining Variability Within the Volcanism-Climate System," *Quaternary Science Reviews*, 19, 417–438. [886]

Zorita, E., Gonzalez-Rouco, F., and Legutke, S. (2003), "Testing the Mann et al. (1998) Approach to Paleoclimate Reconstructions in the Context of a 1000-yr Control Simulation With the ECHO-G Coupled Climate Model," *Journal of Climate*, 16, 1378–1390. [885]

Comment: Hierarchical Statistical Modeling for Paleoclimate Reconstruction

Noel CRESSIE and Martin P. TINGLEY

The article by Bo Li, Douglas W. Nychka, and Caspar M. Ammann (hereafter, LNA) has several goals. It considers the important problem of reconstruction of past (over a period of more than 1000 Years Before Present) climate from multiproxy data, and it directly recognizes the various uncertainties in this undertaking. These uncertainties are expressed through (conditional) probability distributions in a framework known to readers of this journal as hierarchical statistical modeling. LNA use a physical–statistical model that also includes climate forcings, and their statistical inference is Bayesian. Rather than using actual multiproxy data, LNA simulate their data. Then they design a computer-simulation experiment to assess the value of including the various (simulated) proxies and the forcings. The design of the experiment, its analysis, and the conclusions obtained from it, are intended to guide climate scientists towards more precise inferences when carrying out actual paleoclimate reconstructions. Our discussion of LNA in the sections that follow considers both the scientific and statistical goals summarized above.

1. INTRODUCTION

Because LNA use pseudo-proxy data, not real data, we shall examine the way these pseudo-proxy data were created. Based on a combination of scientific expertise and statistical intuition, time series that mimic tree-ring, borehole, and pollen data were synthetically produced from the output of a general circulation model (GCM) of the climate system. This is akin to the biochemist conducting experiments on lab animals, with the goal being to eventually take the results from laboratory to bedside (a goal of Transformative Medicine). LNA realize the importance of calibrating their proxy data (the "lab animals") to the real data (whose analogue would be the "patients").

LNA use a methodology we call here posterior analysis (whose analogue might be the "treatment"), that may be new to the paleoclimate-reconstruction community, but it is well known to statisticians. Posterior analysis resulting from hierarchical statistical modeling is a powerful way to account for uncertainties in all aspects of a scientific study. The main strength of a hierarchical model (HM) is also a point of difficulty, namely that all these uncertainties have to be expressed through (parametric) probability distributions. This might not be easy for a paleoclimate scientist to do, and hence the statistician's involvement is needed in a posterior analysis from the "get-go." As part of our discussion, we shall examine the appropriateness of the HMs proposed by LNA.

LNA use a computer simulation experiment (i.e., try it on the lab animals first!) to determine the worthiness of a Bayesian HM to address this highly complex climate-reconstruction problem. Their experiment should be assessed like any other, in terms of the basic principles of blocking, randomization, and replication (Fisher 1935), and in terms of the responses that are studied to answer the questions that provoked the experiment. In LNA's analyses, the responses all depend on the posterior distribution obtained for the various HMs that were fitted, which is consistent with their (Bayesian) hierarchical modeling approach.

Obtaining posterior distributions from an HM usually requires a considerable investment in computation. Sometimes, modeling decisions in the HM are made more for computational reasons than scientific ones. All of us who use the HM approach are faced with these compromises, and we discuss this in the context of LNA's analyses. In the sections that follow, we expand on all of these issues raised in our introduction.

Noel Cressie is Director of the Program in Spatial Statistics and Environmental Statistics, Professor of Statistics, and Distinguished Professor of Mathematical and Physical Sciences, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247 (E-mail: *ncressie@stat.osu.edu*). Martin P. Tingley is Post-Doctoral Fellow, Statistical and Applied Mathematical Sciences Institute (SAMSI), P.O. Box 14006, Research Triangle Park, NC 27709-4006 (E-mail: *mtingley@samsi.info*). This research was carried out while Cressie was visiting SAMSI under the 2009–2010 Program, "Space–Time Analysis of Environmental Mapping, Epidemiology and Climate Change." It is supported by the National Science Foundation under agreement DMS-0635449. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

© 2010 American Statistical Association
Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap10318

2. DESIGN OF THE COMPUTER SIMULATION EXPERIMENT

LNA use a General Circulation Model (GCM) as the basis of an experiment to see whether an HM approach to paleoclimate reconstruction of temperature is worthwhile. A large part of their article discusses the design of the experiment, and we start with that.

Experimental design has its foundations set out in the book by Fisher (1935); the three basic tenets are blocking, randomization, and replication, and they are by now well accepted by scientists. Think of the "treatment" in this experiment of LNA's as the generic posterior analysis using Bayes' Theorem and MCMC. Then the "experimental units" are the various HMs outlined in LNA's Section 4.3. It is now noticeable that their experiment involves only *one* treatment. It is unusual to do an experiment without another treatment to compare to; in this case, it might be a standard analysis in the paleoclimate-reconstruction literature, such as the RegEM method of Schneider (2001). Even if LNA's posterior analysis does well, does it do better than RegEM, say, or any other method a paleoclimate scientist might use (see, e.g., Jones et al. 2009)? Science advances by replacing an inferior methodology with a superior one, whose inferiority is ideally established through a designed experiment.

LNA *do* carry out "blocking"; see their Section 4.3 where the different factors are listed. These correspond to the various combinations of data and terms included in the HM (e.g., temperature process model without external forcings, the "oracle" proxy, etc.). However, all their blocks are of size one, because they only consider one treatment.

There is a component of "randomization" in the experiment, but not in the sense that Fisher meant it. Fisher was concerned with which experimental unit received which treatment within a block. Here the blocks are of size one, but what *should* happen if two methodologies (e.g., posterior analysis and RegEM) were applied to the (proxy) data and compared? In a simulation experiment, the statistician is able to create two (or more) *identical* experimental units, something a crop scientist could only dream about. (The real-world analogy would be to have homogeneous material—such as a water sample from a lake—that is divided into two parts and a different treatment would be applied to each.) Therefore, in a simulation experiment, randomization of treatment assignment to experimental unit may not be important, depending on the computing resources needed to apply a treatment to an experimental unit.

Finally, how much "replication" do LNA have in their experiment? The experiment has a lot of factors and there are many ways the responses are quantified (see their Section 5), but their experiment has *no* replication. In effect, their study is on only one "lab animal." It is true that they tried to choose a "typical lab animal" by using a GCM simulation of the Earth's climate. However, there are many decisions that go into such climate simulations: A way to introduce replication into this experiment would be to look at an ensemble of such Earth-climate simulators and run several (chosen randomly or purposively) to guard against any objection that the conclusions from this experiment are particular to the climate simulator used. Christiansen, Schmith, and Thejll (2009) make the same point in a recent article published in the *Journal of Climate*.

Statistical simulation experiments should be designed in the same way agricultural, industrial, computer, etc. experiments are designed. Ultimately, the experimenter is looking to attribute the total variability of the responses to various sources in the experiment. Aldworth and Cressie (1999) discuss how this can be done in a systematic way, which they illustrate with a simulation experiment to compare various spatial sampling schemes of an ecological resource. In the case of LNA, with one treatment and one replicate, their analysis can only attribute the total variability to the various factors (or "blocks").

3. MULTIPROXY DATA, REAL AND SIMULATED

In LNA's experiment, the target quantity is the Northern Hemisphere (NH) average temperature, and they build an HM to capture the characteristics of three classes of real-world climate proxies. In order to test their HM, they simulate pseudo-proxy time series using the output of a GCM.

Tree ring pseudo-proxies are constructed by taking the output of the model at a number of grid locations, adding noise, and then removing the 11-year running mean. This construction amounts to high-band-pass filtering the GCM model output that has been purposely noise degraded. Consequently, a tree-ring pseudo-proxy observation at year t is a function of the model output at the corresponding location for years $t - 5$ to $t + 5$. Perhaps a running mean that looks *back* 10 or 11 years would have been a better choice, since a tree ring cannot contain information about the future climate.

We interpret the choices they made for tree-ring pseudo-proxy construction as an attempt to mimic the preprocessing that is often applied to tree-ring data. A number of tree-ring series (each perhaps covering a different time interval) are generally combined to arrive at a single, long, climate-sensitive series, and techniques such as Regional Curve Standardization (e.g., Briffa et al. 1992; Esper, Cook, and Schweingruber 2002) are used to remove biological growth effects from raw tree observations. These steps may result in a tree-ring observation at year t being dependent on local climate for years both before and after t. As referenced in LNA, there is evidence that tree-ring proxies only record faithfully high-frequency climatic changes, due either to this processing or to biology. Some reconstructions (e.g., Moberg et al. 2005) have used tree rings only to infer the high-frequency component of the spatial average temperature series. In their supplementary material, LNA claim that the tree-ring pseudo-proxy construction results in time series that "look similar" to actual (i.e., not band-pass-filtered) tree-ring time series. Our spectral analysis (not shown here) suggests that this is not the case—the actual tree-ring series provided in the supplementary material have an abundance of power at low frequencies (the spectra are red), while the LNA construction results in spectra with a sharp drop-off in power for periods longer than 10 years (as expected).

More realistic tree-ring pseudo-proxies could be created by using the GCM to drive a forward model of tree growth, such as that proposed by Shashkin and Vaganov (e.g., Shashkin and Vaganov 1993; Evans et al. 2006). As a result, the pseudo-proxy time series would approximate more closely the actual proxy time series (i.e., the makeup of the "lab animal" would be closer to that of the "patient").

Pollen pseudo-proxies are created by averaging the model output over a number of $7.5° \times 7.5°$ regions, adding noise, calculating an 11-year running mean, and then sampling every 30 years. This reflects the fact that pollen assemblages record information about the climate over large spatial and long tem-

poral scales. For example, the species composition of a forest stand responds gradually to changes in the climate, while the pollen produced by that stand can travel considerable distances. The implication that an observation of the proxy for a given year contains information about the climate in both the past and the future is perhaps more justified in the case of pollen proxies, which are measured by analyzing small segments of sediment, often from lake-floor cores. Various physical and biological mechanisms can mix the pollen deposited over a number of years, while the measurement process itself could involve sediment accumulated over more than one year.

Borehole pseudo-proxies are formed from spatial averages over $20° \times 20°$ regions by application of the POM-SAT model, which describes the diffusion of surface-temperature perturbations through the bedrock. LNA use POM-SAT to simulate borehole temperature profiles down to 500 m, and then they sample this depth profile every 5 m. The details of the POM-SAT model are not provided.

LNA assume that a borehole profile provides information about surface temperatures at large spatial scales. As a temperature anomaly needs to propagate through rock, to where the measurement takes place, the spatial scale of the information might be considered similar to the depth scale. If this were so, we find the choice of $20° \times 20°$ regions to be too large. By forming the borehole pseudo-proxies from such large regions, LNA likely overemphasize the information in actual borehole proxies for inferring the NH (spatial mean) temperature time series. Their borehole pseudo-proxy construction was probably motivated by the observation that surface temperatures averaged over longer time scales tend to reflect larger spatial scales. A fully Bayesian reconstruction of temperature in the San Rafael region of Utah from borehole proxies is given by Brynjarsdóttir and Berliner (2010).

A better approach would involve constructing the borehole pseudo-proxies from local GCM output, and then using an HM with a spatial component to reconstruct the temperature field through time, not just the NH spatial average. Ideally, this would be calibrated to the heat equation that governs the temporal evolution of subsurface temperatures for which surface temperatures are boundary conditions. The data level of such a hierarchical model could represent the borehole data as reflecting local temperature, smoothed through time, while the process level could model a temperature field that becomes increasingly smooth in space as temporal smoothing increases. We say more about the introduction of a spatial component in Section 6.

For all three classes of pseudo-proxies, LNA add noise to the local GCM temperatures before forming the pseudo-proxies. In other words, additive white noise, as well as the GCM output, are subject to the transforms that create the pseudo-proxies. This decision is motivated by the observation that actual pollen and borehole proxies appear somewhat "smooth" in time. The notion that borehole proxies are temporally smooth is reflected in the data level, as the noise term is also subject to the transformation $\mathbf{M}_B$ [LNA's (4.3)]. This is not the case for the pollen pseudo-proxies, which are modeled at the data level as being subject to additive AR(2) noise [LNA's (4.2)]. The tree-ring pseudo-proxies are likewise modeled at the data level as being subject to additive AR(2) noise that is not filtered by the transform matrix $\mathbf{M}_D$ [LNA's (4.1)]. In short, we see something of a disconnect between the observed properties of proxies, pseudo-proxy constructions, and the assumptions made at the data level of the HM.

LNA investigate the potential of borehole, pollen, and tree-ring proxies, along with estimated time series of forcings, to reconstruct NH mean temperatures. However, as they state clearly, their methodology has not been tested on actual data. In their experiment, the parameters used to transform the GCM output into pseudo-proxies are assumed known—the matrices $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$ that appear in the data level of the model are used to construct the pseudo-proxies. In real paleoclimate reconstructions, this will not be the case. For example, the assumption that a pollen observation reflects some weighted average of temperatures over a number of years is likely reasonable, but the *number* of years reflected in that observation, and the *weights* associated with the averaging, will in general *not be known*.

LNA assume that each pseudo-proxy time series has a (different) linear relationship with the transformed unknown temperatures $\mathbf{T}$; for each time series, they infer two regression parameters, and for each proxy type they infer three parameters [two AR(2) coefficients and a variance] for the error process. For 15 tree-ring time series, this amounts to $2 \cdot 15 + 3 = 33$ parameters.

Another strategy would be to infer the averaging weights in the matrices $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$, in which case the scaling coefficients ($\beta_{i,D}$, $\beta_{j,P}$, and $\beta_{k,B}$) in the data model are redundant. We shall now investigate the consequences of this for tree-ring proxies. If the matrix $\mathbf{M}_D$ is assumed to represent a stationary linear transform of the temperatures $\mathbf{T}$ within plus or minus five years of a given observation, then an additional 11 parameters must be inferred. If there are 15 tree-ring proxies, each with an intercept only, and assumed to have common error process parameters, then there are $11 + 1 \cdot 15 + 3 = 29$ parameters to be estimated for the tree-ring proxies. As the matrices $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$ are not known in real-world applications, this (slightly more parsimonious) model may be more realistic. The impacts of these differing modeling choices would certainly be of interest to the scientific community and could be included in a future simulation experiment.

In this paper, LNA have tested the ability of a BHM to reconstruct past temperatures, given *pseudo-proxies* obtained by applying different, *known* transforms to the climate-model output. While these transforms are constructed to reflect aspects of tree-ring, pollen, and borehole proxies, they are at best simple approximations to the processes that generate the actual proxies. In addition, assuming that these transforms are known eliminates a source of uncertainty which, in real-world applications, could be large. To sum up our discussion in this section, we see many ways that the "lab animal" is different from the "patient."

4. THE HM USED FOR RECONSTRUCTING PAST CLIMATE

4.1 Hierarchical Modeling Choices

There are several choices made by LNA when building the HM in their Section 4.1. First, they introduce forcings $\mathbf{S}$ (solar irradiance), $\mathbf{V}$ (volcanism), and $\mathbf{C}$ (greenhouse gases, represented by the concentration of CO_2) into the model. Strictly speaking, the forcings should appear in the process model (4.5) in terms of their "noise-free" versions, $\mathbf{S}_0$, $\mathbf{V}_0$, and $\mathbf{C}_0$. In terms of the probability structure defined by LNA's HM, all distribu-

tions are in fact *conditional* on $\mathbf{S}$ and $\mathbf{C}$. This is an assumption that we invite LNA to comment on. Was the reason for this a pragmatic one that kept the number of unknowns to a manageable size? Notice that a data model for $\mathbf{S}$ and $\mathbf{C}$ would introduce unknowns $\mathbf{S}_0$ and $\mathbf{C}_0$ into the HM and hence into the MCMC.

There was another hierarchical modeling decision made by LNA that we would like to invite comment on. While it is not explicitly stated, LNA assume that the observed instrumental temperatures $\mathbf{T}_2$ have *no* measurement error; however, this does not seem to reflect paleoclimate scientists' understanding (e.g., Brohan et al. 2006). Therefore, we suggest that the data stage should include one more equation:

$$\mathbf{T}_2 = \mathbf{T}_{2,0} + \boldsymbol{\varepsilon}_2.$$

Then, at the process-stage, their (4.5) should be in terms of $(\mathbf{T}_1', \mathbf{T}_{2,0}')$. In fact, we find the notation, $\mathbf{T}_1$, for past temperature misleading; in line with other notation, we suggest that it be replaced with $\mathbf{T}_{1,0}$. Consequently, when LNA state [just after (4.5)], "The target is to estimate $\mathbf{T}_1$ given $\mathbf{T}_2$, the proxies and the forcings," we suggest that their target should be to make inference on the *unknown* $\mathbf{T}_{1,0}$, given the temperature *data* $\mathbf{T}_2$. Moreover, we would extend this to making inference on the unknown $\mathbf{T}_{2,0}$ as well.

It is clear that LNA have assumed $\mathrm{var}(\boldsymbol{\varepsilon}_2) = \mathbf{0}$, without explicitly stating it. Even if it were true, we find that it helps to distinguish between the (potentially) observed temperatures, $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')'$, and the unknown true temperatures, $\mathbf{T}_0 \equiv (\mathbf{T}_{1,0}', \mathbf{T}_{2,0}')'$. Then (4.1), (4.2), and (4.3) should be formulated conditional on $\mathbf{T}_0$, and the focus of the study would be on gaining knowledge about the unobserved past temperatures, $\mathbf{T}_{1,0}$ (and $\mathbf{T}_{2,0}$ as well), from all relevant data sources (including $\mathbf{T}_2$).

4.2 Spatial Sampling Issues

LNA provide justification for the spatial distribution of the pseudo-proxies they chose (Figure 2 of LNA) for only the boreholes, saying in this case that the "distribution of those locations reasonably represents the spread of real borehole data." Now consider the particular spatial distribution they chose for the tree-ring pseudo-proxies (Figure 2 of LNA). First, a number of the tree-ring pseudo-proxies are located in the Southern Hemisphere, despite their goal of wanting to reconstruct NH mean temperatures. Second, LNA locate several tree-ring pseudo proxies in the tropics, despite the fact that trees in the tropics do not generally develop annual rings, due to the the lack of strong seasonality. Third, LNA locate a tree-ring pseudo-proxy in Eastern Greenland, north of 75°N, and certainly north of the tree line.

More generally, the spatial distribution of the pseudo-proxies presumably impacts their ability to infer the NH spatial average. All things being equal, we might think that a regularly spaced distribution of locations would be preferred (see, e.g., Aldworth and Cressie 1999). However, the surface-temperature field is inhomogeneous and the spatial distributions of tree-ring and pollen proxies used in any actual application are limited to particular geographical areas. Indeed, the locations of proxies in published reconstructions (e.g., Mann et al. 2008) could have been used to inform the locations of the pseudo-proxies, but this was not a factor controlled in LNA's experiments. We suggest that in a future simulation experiment, the opportunity should be taken to consider the effect of spatial locations of proxies on inference for NH mean temperatures.

4.3 Accounting for Nonlinearities in the HM

We have already noted in Section 3 that each pseudo-proxy time series is *linearly* related to the true temperature time series. There is growing evidence that some tree-ring proxies, particularly those at high northern latitudes, have become less sensitive to changes in local temperature over the last few decades (Briffa et al. 1998; Jones et al. 2009). This so-called "divergence" problem could be explained by nonstationarities or nonlinearities in the tree-ring temperature relationship, or by the presence of confounding covariates that are generally not included in paleoclimate reconstructions. The problem with capturing nonlinearities in the data stage of an HM (as do LNA) is that the nonlinearities are assumed part of the measurement error and filtered out by the posterior analysis.

We would like to finish this section by augmenting our discussion of LNA's Equation (4.5). They assume that the forcings in (4.5) are additive in $\mathbf{S}$ (or $\mathbf{S}_0$), $\mathbf{V}_0$, and $\mathbf{C}$ (or $\mathbf{C}_0$), an assumption that seems to be supported by the IPCC's Fourth Assessment report (Forster et al. 2007). However, this only refers to the lack of interaction between the $\mathbf{S}$, $\mathbf{V}_0$, and $\mathbf{C}$. Now, the radiative forcing associated with CO_2 increases as the log of the mixing ratio (Forster et al. 2007). Furthermore, it seems clear from LNA's Figure 1 and the multiplicative measurement error in their (4.4), that $\mathbf{V}_0$ should also be expressed on the log scale. Therefore, we suggest that Equation (4.5) be modified to be additive in $\mathbf{S}$, $\log(\mathbf{V}_0)$, and $\log(\mathbf{C})$, where the log of a vector is interpreted as the vector of elementwise logs.

5. INFERENCE IN THE PRESENCE OF UNCERTAINTY: WHAT ARE THE QUESTIONS AND HOW ARE THEY ANSWERED?

LNA's scientific goal is to investigate the value of including proxies with different spatial and temporal relationships, as well as various forcings, into a paleoclimate reconstruction based on posterior analysis. Their statistical goal is to assess a computer simulation experiment designed around how an HM includes aditional data for paleoclimate reconstruction.

Their posterior analyses are done carefully but, as discussed in Section 2, we believe the design lacks a competing methodology and there should be some replication. Perhaps the lack of replication is an explanation for the questions below. In Figure 5, under the factor combination (Noise, T1, D), the "forcings" bias is extemely negative, resulting in a worse *rmse* when forcings are included. Does this make sense? Also, does the borehole proxy really lead to a less-biased reconstruction? And shouldn't the "oracle" be best in terms of virtually any skill measure? (It's not; see their supplementary material.)

A lot of effort was put into understanding which factors in the reconstruction are important; the end product is a comparison of a number of possible reconstructions, based on the bias and mean squared prediction error of posterior means. (We believe that LNA used *posterior means*, but we could not actually find where they specified which posterior summary was used to define the reconstructions.) Coverage rates of posterior credible intervals are investigated in LNA's supplementary material and,

there, the performance of even the "oracle" is quite poor. (The coverage rate is an attractive measure of skill since it is unitless and intuitively interpretable.) For example, with a nominal coverage rate of 90%, the oracle proxy only gives 65% coverage! Presumably, this difference is an indication of the inherent limitations involved in inferring the NH mean using proxies at the particular locations used by LNA.

Is this experiment applicable to real-world proxies? We have already mentioned that the matrices $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$, given in their data stage, are not known in practice. Either estimating them, putting a prior on them, or carrying out a sensitivity analysis should be done before applying the HM proposed by LNA to real-world data.

LNA investigate the interaction of proxy and climate-forcing information in their roles of reconstructing climate. They investigate the impacts of including different combinations of proxies, of including noise in the proxy construction, including the forcing time series, and of modeling the temperature process over the entire time span or only over the reconstructed time interval. They conclude that it is important to include information about the target time series at a wide array of frequencies. If tree rings only reflect high-frequency climate variability (as is assumed by LNA), then including the different forcing time series improves results. Including proxies that reflect the lower-frequency variability of the target times series can partially replace the role of the forcing time series. These are nice, "take-home" conclusions, and they extend the results of Moberg et al. (2005) to include the effects of climate forcings. A consistent message seems to be that having information on different time scales is essential for arriving at skillful reconstructions of past climate.

According to the process stage of the model [Equation (4.5) from LNA] the NH mean temperature time series is a linear combination of the three forcing time series plus AR(2) noise. It would be interesting to re-run the analysis using different combinations of the forcing series, to investigate, for example, the impact of including solar variability in the model. (Climate change skeptics often attribute temperature changes to solar forcings, in place of the more common attribution given to greenhouse gases.) This would address the influences of the various forcing time series, similar to the way in which LNA address the influence of the various types of pseudo-proxies.

There are several other factors that could be investigated, which we have discussed earlier but group together here: Are the results robust to different runs of the climate model, or to output from different models? The spatial network of proxies is held fixed across all experiments, although the spatial distribution of these series must have an impact on their ability to infer NH mean temperature. How variable are the results as a function of the spatial network? What is the optimal spatial design, and is this dependent on the GCM output or the particular GCM chosen? Finally, LNA discuss the possibility of dating errors when dealing with actual proxy time series, particularly for pollen observations. Sensitivity to mild dating errors could have been explored in their experiment.

6. SPATIAL MODELING FOR PALEOCLIMATE RECONSTRUCTION

The spatial aspect has not been featured in LNA's data stage or process stage, something we would like to discuss in this section. We have suggested above that $\mathbf{T}_0$ is the appropriate time series of temperature, where the spatial component has been averaged out over the NH. Now think of a time series of *spatial* temperature processes, which we write as $\{\mathbf{T}_{0,1}(\mathbf{s}) : \mathbf{s} \in globe\}$, $\{\mathbf{T}_{0,2}(\mathbf{s}) : \mathbf{s} \in globe\}, \ldots,$ the present-day temperature process over the globe. Write this time series of spatial processes as $\mathbf{T}_0(\cdot)$. Then the spatial component might be introduced into the data stage through spatially varying parameters in (4.1), (4.2), and (4.3), including parameters found in $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$ (see our Section 3). The current LNA model relates each proxy observation to a number of years of the NH mean times series (via the matrices $\mathbf{M}_D$, $\mathbf{M}_P$, and $\mathbf{M}_B$). A spatial adaptation of $\mathbf{M}_D$ would reflect the local (in space) temperature value, whereas that of $\mathbf{M}_P$ would reflect regional temperature values. As discussed above, it is our view that $\mathbf{M}_B$ should reflect the local temperature value, but with a space–time covariance that models an increasing spatial range with longer temporal averaging. Clearly, the curse of dimensionality needs to be taken into account when including spatial dependence in the model, which we discuss below.

At the process stage, the forcings likely mix well enough over annual time scales that we do not have to include any spatial variability in that part of the model. However, the error term $\boldsymbol{\varepsilon}_T$, which accounts for process variability in LNA's key regression Equation (4.5), should now be spatio-temporal with nonstationary spatial covariances.

With all this extra structure, any posterior analysis runs the risk of being overwhelmed by high dimensionality. One way to reduce the dimensionality is to use a spatio-temporal random effects (STRE) model, as in Cressie, Shi, and Kang (2010), where a spatio-temporal analysis was done on a very large remote-sensing dataset. In LNA's terminology, write the (now) spatio-temporal error $\boldsymbol{\varepsilon}_T$ in their (4.5) as $\boldsymbol{\varepsilon}_T \equiv (\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ldots, \boldsymbol{\varepsilon}_t', \ldots)'$, and assume

$$\boldsymbol{\varepsilon}_t \equiv S_t \boldsymbol{\eta}_t + \boldsymbol{\xi}_t; \qquad t = 1, 2, \ldots, T,$$

where the matrices $\{S_t : t = 1, 2, \ldots, T\}$ are made up of known spatial basis functions and $\{\boldsymbol{\eta}_t : t = 1, 2, \ldots, T\}$ is an r-dimensional *vector autoregressive* time series. Importantly, r is fixed. In Cressie, Shi, and Kang (2010), r was on the order of 100. The last term in the STRE model, $\{\boldsymbol{\xi}_t : t = 1, 2, \ldots\}$, captures fine-spatial-scale variability.

The STRE model inherits nonseparable, nonstationary spatio-temporal covariances, an attractive feature since stationarity is not expected over a global spatial scale and a millennial temporal scale. Critically, the dimension reduction allows very fast matrix inversions in an MCMC. Tingley and Huybers (2010a, 2010b) present a spatio-temporal BHM for paleoclimate reconstructions that assumes separability of the spatial and temporal variability in order to achieve a computationally feasible MCMC. Both approaches use a sequential updating procedure to speed up inferences and they are $O(T)$ in computational complexity. The dimension reduction appears to be needed when spatial-data sizes go beyond about 2000 observations per time point.

Including a spatial model will produce estimates of the spatial mean and its associated uncertainty that are consistent across global, hemispheric (including the NH), continental, and regional scales. Indeed, results included in LNA's supplementary material point to the limitations of inferring a spatial average without modeling the spatial covariance. LNA compare

reconstructions based on the oracle proxies to those based on dendro and pollen proxies using rank-verification histograms. They note that the shapes of the rank-verification histograms are the same for each, but that the time series of temperature observations at the particular set of locations they have chosen cannot capture all aspects of the spatial average temperature.

7. CONCLUSIONS

LNA have presented an HM for reconstructing past climate from various types of (pseudo-)proxy and forcing information, which represents a 21st-century statistical approach to paleoclimate reconstruction. One of the great advantages of an HM is the conceptual ease with which different forms of uncertainty can be included, as well as the transparency of the physical and statistical modeling assumptions. While we feel that there are a number of aspects of LNA's HM that could be improved upon, their efforts do represent a substantial step forward for the paleoclimate-reconstruction community, whose statistical approaches are summarized in NRC (2006). The suggestions we have made are in support of posterior analysis from an HM approach, and LNA's paper indicates that posterior analysis on actual paleoclimate data will advance our understanding of the Earth's past climate (as well as quantify the associated uncertainties). We look forward to such efforts appearing in the literature in the near future.

ADDITIONAL REFERENCES

Aldworth, W. J., and Cressie, N. (1999), "Sampling Designs and Prediction Methods for Gaussian Spatial Processes," in *Multivariate Analysis, Design of Experiments, and Survey Sampling*, ed. S. Ghosh, New York: Marcel Dekker, pp. 1–54. [896,898]

Briffa, K. R., Jones, P. D., Bartholin, T. S., Eckstein, D., Schweingruber, F. H., Karlen, W., Zetterberg, P., and Eronen, M. (1992), "Fennoscandian Summers From AD 500: Temperature Changes on Short and Long Timescales," *Climate Dynamics*, 7, 111–119. [896]

Briffa, K. R., Schweingruber, F. H., Jones, P. D., Osborn, T. J., Shiyatov, S. G., and Vaganov, E. A. (1998), "Reduced Sensitivity of Recent Tree-Growth to Temperature at High Northern Latitudes," *Nature*, 391, 678–682. [898]

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*, 2, 99–113. [898]

Brynjarsdóttir, J., and Berliner, L. M. (2010), "Bayesian Hierarchical Modeling for Temperature Reconstruction From Geothermal Data," Technical Report 842, The Ohio State University, Dept. of Statistics, Columbus, OH. [897]

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. [896]

Cressie, N., Shi, T., and Kang, E. L. (2010), "Fixed Rank Filtering for Spatio-Temporal Data," *Journal of Computational and Graphical Statistics*, 19, 724–745. [899]

Evans, M., Reichert, B., Kaplan, A., Anchukaitis, K., Vaganov, E., Hughes, M., and Cane, M. (2006), "A Forward Modeling Approach to Paleoclimatic Interpretation of Tree-Ring Data," *Journal of Geophysical Research*, 111, G03008, DOI: *10.1029/2006JG000166*. [896]

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh, U.K.: Oliver & Boyd. [895,896]

Forster, P., Ramaswamy, V., Artaxo, P., Berntsen, T., Betts, R., Fahey, D. W., Haywood, J., Lean, J., Lowe, D. C., Myhre, G., Nganga, J., Prinn, R., Raga, G., Schulz, M., and Van Dorland, R. (2007), "Changes in Atmospheric Constituents and in Radiative Forcing," in *Climate Change 2007: The Physical Science Basis*, eds. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Cambridge, U.K.: Cambridge University Press. [898]

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. et al. (2009), "High-Resolution Paleoclimatology of the Last Millennium: A Review of Current Status and Future Prospects," *The Holocene*, 19, 3–49. [896,898]

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. [896]

Shashkin, A. V., and Vaganov, E. A. (1993), "Simulation Model of Climatically Determined Variability of Conifer's Annual Increment (on the Example of Common Pine in the Steppe Zone)," *Russian Journal of Ecology*, 24, 275–280. [896]

Tingley, M. P., and Huybers, P. (2010a), "A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and Applications to Paleoclimate Reconstruction Problems," *Journal of Climate*, 23, 2759–2781. [899]

——— (2010b), "A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 2: Comparison With the Regularized Expectation–Maximization Algorithm," *Journal of Climate*, 23, 2782–2800. [899]

Comment

Eugene WAHL, Christian SCHOELZEL, John WILLIAMS, and Seyitriza TIGREK

1. INTRODUCTION

This is a wonderful time to work on the reconstruction of past climates. For periods prior to the modern era of instrumental climate data we are reliant on quantitative "proxy" information to indirectly "reconstruct" climatic conditions, and in some situations nonquantitative historical documentation is also useful for this purpose. Proxy-based climatic information has been shown to be remarkably effective at capturing longer-term and shorter-term climate information in preinstrumental times (pre ∼1850–1950, depending on the climate variable and location) in a wide variety of situations, and has led to crucial breakthroughs in understanding earth system history and dynamics. A longer-term example is the documentation of glacial-interglacial cycles over the past ∼2.75 million years, and linking these to variations in earth's orbital eccentricity, obliquity (tilt), and precession (when orbital extremes occur during the cycle of seasons) (Im-

Eugene Wahl is Physical Scientist, Paleoclimatology Branch/World Data Center for Paleoclimatology, National Climatic Data Center (NOAA-NCDC), National Oceanic and Atmospheric Administration, Boulder, CO 80305 (E-mail: *Eugene.R.Wahl@noaa.gov*). Christian Schoelzel is Meteorologist, Meteorological Institute, University of Bonn, 53121 Bonn, Germany (E-mail: *schoelzy@gmx.net*). John Williams is Bryson Professor of Climate, People, and Environment, Department of Geography, University of Wisconsin, Madison, WI 53706 (E-mail: *jww@geography.wisc.edu*). Seyitriza Tigrek is Doctoral Student, Electrical Engineering, University of Colorado, Boulder, CO 80309 (E-mail: *Seyitriza.Tigrek@Colorado.edu*). Support for the research reported in the Supplemental Online Material was provided by NOAA and through NSF CMG Collaborative Research award 0724619. The authors would like to thank Dr. Juerg Luterbacher, four other reviewers, and the editor for suggestions that significantly improved the text.

© 2010 American Statistical Association
Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap10295

brie and Imbrie 1979). A shorter-term example is the recognition that very long and intense droughts have occurred in North America and eastern Asia (as examples—these are the regions with the most complete paleodrought coverage), which have lasted far longer than the worst droughts in the instrumental record that have been used for planning in water management applications (e.g., Cook et al. 2004). Paleoclimate reconstruction is currently in a time of explosion in both interest in its results and the development and testing of new methods. The work by Li, Nychka, and Ammann (this issue) to study the capabilities of Bayesian Hierarchical Modeling for use in paleoclimate reconstruction is an important "new shoot" of this development.

2. TECHNICAL BACKGROUND

In many applications, proxy information can be calibrated against a climatic variable (or variables) of interest. In some cases, such calibrations can be physically deterministic, derived from first principles and/or carefully controlled laboratory experiments. More typically, they are purely statistical, developed over a period of common overlap between the climate and proxy data, and then the calibration relationship is applied to yield quantitative estimates of the climate variable as far back in time as the proxy data allow. Both kinds of calibration assume stationarity of the relationships modeled over time, as does the Bayesian hierarchical approach evaluated by Li, Nychka, and Ammann. Statistical calibrations have involved a variety of simple to more complex mathematical models, often utilizing one of a number of regression-based approaches. Two of the most common are: (1) direct (or indirect) regression of the proxies against, for example, temperature, in either simple or multiple regression forms (cf. Frank et al. 2010); and (2) a more complex process in which one of these closed-form regression approaches (or an iterative procedure such as regularized expectations maximization, RegEM) is used fit the proxy data to the component time series weights of an orthogonal spatial decomposition of the target instrumental data field, and the resulting fitted component time series are then input back into the appropriate eigen/singular value (SV) expansion to yield estimated spatial fields (e.g., Luterbacher et al. 2004; Mann et al. 2007, 2009). These methods have proved very effective in making *expected value* reconstructions of climate at both single sites and for entire climate fields up to global scale. Simultaneous whole-field reconstructions, in particular, have been one of the key successes of these approaches, because they provide site-to-site-consistent spatial information about a dynamical system (climate) that is inherently spatial in its character. (In this context, it is worth keeping in mind that even with modern instrumentation, including satellites, providing systematic whole-field climate data coverage remains difficult.) In addition, a variety of methods for validating reconstruction results have been derived from mathematical statistics, including econometrics and the theory of calibration and validation (cf. Cook, Briffa, and Jones 1994; Wahl and Ammann 2007). Additional work is currently being done in this area to refine and extend the efficacy of validation methods. Much work is also underway to consider the impact of, and solutions for, classically known problems in regression such as errors-in-variables (EIV), heteroscedasticity, multicollinearity, etc. (e.g., Hegerl et al. 2007; Mann et al. 2007; cf. Ammann, Genton, and Li 2010 both for new methodology and for a good review of the relevant literature). For example, one reason (among several) for using eigen/SV decompositions is to help limit stochastic noise in the instrumental and proxy data by truncating the eigenvectors/EOFs used in the reconstruction process, in particular to help reduce EIV issues when the form of the regressions is climate $= f$(proxy).

3. USE OF BAYESIAN METHODS

Perhaps most importantly at the current time, much work is being done to develop appropriate models that can be used to generate reconstruction *ensembles*, for the purpose of characterizing reconstruction *uncertainty*. This work includes (but is not limited to): (1) identifying stochastic models for this purpose and then applying these via Monte Carlo methods to make random draws from the estimated distribution of reconstructions (Li, Nychka, and Ammann 2007; Supplemental Online Material here); and (2) a more "engineering" style approach that describes reasonable ranges of free analyst choices in model specification per se, and then uses all possible models within these ranges to generate the reconstruction ensemble from which draws can be taken (Frank et al. 2010). It is, of course, possible to combine these approaches to evaluate the combined effects of stochastic and model selection uncertainties. The work of Li, Nychka, and Ammann in this issue is part of a fundamentally new approach to the problem of generating statistically appropriate ensemble distributions of reconstructions from which random draws can be taken (e.g., Haslett et al. 2006). In this approach, a clearly specified hierarchy of mathematical models is formulated as conditional probability density functions, which can then be combined via Bayes' Theorem to yield an estimated joint posterior distribution of the climate variable(s) being reconstructed, the proxies used in the process, and the parameters in the hierarchical model. The hierarchy of models in this formulation has three stages: (1) a *data stage*, in which the proxy or proxies used in the reconstruction are modeled in a "forward" way as causally dependent on the climate variable(s) of interest; (2) a *process stage*, in which the climate variable(s) is (are) modeled, for example, as a simple autoregressive process or, more appropriately from a physical standpoint (as Li, Nychka, and Ammann do), as causally dependent on known factors that "force" the climate system (such as solar output, volcanic aerosol release, and greenhouse gas concentrations); and (3) the assumed *prior* distributions of the unknown parameter values at both the data and process stages.

Although complex, this Bayesian hierarchical modelling (BHM) approach offers distinct resources in relation to the "traditional" reconstruction approaches mentioned above. First, and most importantly, the specification of the full model in an explicit mathematical/statistical form that leads to direct estimation of the posterior distribution means that full uncertainty estimation is inherent in the reconstruction process from start to finish. It does not have to be developed in a component-by-component manner that runs a risk of becoming, to a lesser or greater degree, ad hoc, or simply not tractable per se. Second, explicit specification of the mathematical reconstruction model used allows each part of the reconstruction process to be formally examined for its physical realism, climatological

validity, etc. This transparency of assumed model(s) is not always as easily considered in the other approaches mentioned. Third, especially when the process stage model relates the climate to its known forcing factors, the use of forward models in the data and process stages is more physically realistic in terms of cause-and-effect relationships than the common specification of physically "inverse" models, for example, of the form climate $= f$(proxy).

Li, Nychka, and Ammann apply the BHM approach in a reconstruction simulation experiment (RSE) context, constructing the data models and deriving climate information (in this case, average Northern Hemisphere surface temperature) from a long paleoclimate run of a complex three-dimensional atmospheric-ocean general circulation model (AOGCM). The RSE approach represents a major step forward in examining paleoclimate reconstruction efficacy, providing an experimental "test bed" in which reconstruction approaches can be evaluated prior to implementation on real-world data. Li, Nychka, and Ammann exploit the RSE context to examine a factorial set of simple to more complex BHM designs. Relaxing the physically realistic process model of temperature driven by climate forcings to a simple mean function allows examination of the usefulness of incorporating forcing data in the reconstruction process, which has only begun to be systematically explored in paleoclimate science (cf. Lee, Zwiers, and Tsao 2008, who examine use of the Kalman filter for this purpose in a seminal RSE study).

Li, Nychka, and Ammann's design additionally allows examination of another important area of development in climate reconstruction: the value of using one proxy type only (preserving homogeneity of climate response characteristics, but with potential frequency sensitivity limitations) versus the combining of proxies with different response characteristics, but with a wider range of frequency sensitivities (e.g., Moberg et al. 2005). They construct simulated proxies with three kinds of frequency sensitivities to enable this examination, each with and without a stochastic noise component: (1) a high-frequency-only proxy (defined as carrying annual-to-decadal resolution), sampled at annual time steps (mimicking a potential worst-case of the information available in tree rings); (2) a mid and low-frequency proxy (defined as carrying decadal-to-lower resolution), sampled at 30-year time steps (mimicking the information typically available from pollen preserved in sedimentary deposits); and (3) a low-frequency-only proxy (defined as carrying multidecadal-to-multicentennial resolution), sampled at 100 five meter depth intervals (mimicking the information available from borehole temperature profiles as a partial function of heat diffusion from the surface, which is increasingly smoothed the farther in the past a surface temperature anomaly occurred). (Note that the relationship of depth to time in the diffusion model is not described by the authors.) Li, Nychka, and Ammann also add what they term an "oracle proxy," which is formed as the "true" model temperature time series at the proxy locations, with and without stochastic noise (allowing examination of the extent to which these locations carry sufficient information to capture the full hemispheric mean). This richness of factorial examination is unique in the RSE literature to date, and is even more powerful when combined with examination of the presence/absence of forcing information.

It should be noted that RSE comparison of reconstruction models, noise levels, proxy richness and spatial coverage, calibration choices, and other factors has been done in this relatively new area of study, although typically only one or a few factors are evaluated in a given study (cf. Mann and Rutherford 2002; Rutherford et al. 2005, 2010; Bürger, Fast, and Cubasch 2006; Ammann and Wahl 2007; Mann et al. 2007; Lee, Zwiers, and Tsao 2008; Jones et al. 2009; Riedwyl et al. 2009; von Storch, Zorita, and González-Rouco 2009; Smerdon et al. 2010). The kind of richer factorial RSE developed by Li, Nychka, and Ammann in a BHM context can also be done with the traditional reconstruction methods described. Large n-way sets of simulated proxy data could be developed that include stochastic noise components, and these could be used in various reconstruction scenarios in Monte Carlo fashion. Forcings information could also be incorporated in a variety of ways. However, carrying over uncertainty estimation into a real-world context would continue to be less "natural" than in the Bayesian context, as described.

4. SIMULATION RESULTS

4.1 Inclusion of Forcing Information

The results from Li, Nychka, and Ammann's examination show several salient features. Most prominently, the inclusion of forcing information generally leads to greatly improved reconstructions, especially in terms of reduction of variance. Forcing information also is greatly valuable to reduce reconstruction bias when the temperature process model is assumed to be constant across both the known, instrumental time period, and the time period to be reconstructed utilizing the proxy data. When this assumption is relaxed (representing a partial relaxation of the stationarity assumption described above) bias is generally unaffected by inclusion of forcing information, with the exception of reconstructions based on noisy tree ring information only, in which case adding forcing information increases bias towards too-low values. Although this relaxation leads to reduced bias (which can be noted particularly in the absence of noise in the proxy data), it is associated with increased variance, which should be taken into account before choosing this variant of the BHM over its stationary complement.

4.2 Inclusion of Different Kinds of Proxy Information

When pollen-type proxy data are included with either tree ring or borehole-type data, or these types combined, the value of forcing information to reduce bias under the full stationarity assumption is significantly lessened. The inclusion of pollen-type information also reduces variance in all cases. These results suggest that the mid-to-low frequency information carried by the pollen data represents a significant portion of the additional information content that the forcing data carry and could, to an extent, substitute for the forcing information. It should be noted that establishing the depth-age relationship in the sedimentary deposits from which fossil pollen is typically extracted depends on dating methods that carry significant inherent imprecision and age-dependent possibilities of bias (such as ^{14}C dating), and can also be affected by offsets in the time of pollen production versus its deposition, similar offsets between production of the material dated and its deposition, nonconstant

rates of sedimentation that make interpolation between dated strata uncertain, and other factors. These sources of error could potentially be modeled in the BHM framework (cf. Haslett and Parnell 2008), but would introduce additional imprecision in the information content of pollen data. Thus, the frequency capacity assumed in Li, Nychka, and Ammann's tests is a best-case situation (approached in the real world at relatively rare sites that produce annual/subannual depositional bands called varves, which, presumably, can be counted like tree rings), and the partial capacity of pollen-derived information to substitute for forcing data in BHM reconstructions likely represents a theoretical optimum that is not generally obtainable in practice.

By construction, tree-ring-type data are necessary to obtain higher-frequency information at subdecadal time scales in Li, Nychka, and Ammann's tests. This kind of dependence is also seen in many real-world paleoclimate reconstruction situations, and the availability/unavailability of dendrochronological information is one of the key features that differentiate paleoclimatology of the past 1–2 millennia from reconstruction work further back in time. That very low frequency information provided by borehole-type data had little effect on reconstruction quality is surprising, and likely is due to the age-related smoothing of borehole information the farther one goes back in time. However, the geographic coverage of boreholes is extensive, and exploiting them for the kind of low frequency information they can offer, whether within a BHM framework or not, continues to be a valuable tool to paleoclimatologists, not least because the inversion relationships that underlie borehole reconstructions are based on well-established physical laws of heat diffusion in solids.

5. LOOKING TO THE FUTURE

The BHM RSE study by Li, Nychka, and Ammann, along with a few empirical studies using Bayesian approaches, demonstrates that *Bayesian methods, and the BHM structure in particular, represent an important new application in the toolkit of paleoclimate reconstruction*. By offering an explicit model structure and systematic treatment of uncertainty from "start to finish," BHM offers both a new methodology and a reference point for work to extend traditional paleoclimate methods to build well-composed reconstruction ensembles. It will be particularly valuable to make carefully designed parallel comparisons between BHM and traditional reconstruction methods extended to generate true ensembles; to determine the relative efficacy of both approaches for understanding past climates, and also to examine potential issues in terms of computational effort and effects that may be caused by relative stability/instability of estimated parameters in the BHM framework. Dealing with dating issues involved with any kind of proxy that is not associated with a physical process with regular time steps (annual growth characteristics in trees, varves in sediments, etc.) is another clear avenue for refinement, as mentioned. Additional dimensions in RSE analyses of BHM performance that would be valuable to incorporate include: (1) joint analysis of more than one climate variable, for example, temperature and precipitation; (2) employing more than one AOGCM run, to enable isolation and elimination of model-specific outcomes; (3) examination of seasonal as well as annual reconstruction performance, as annual averages can mask important seasonal variability; and (4) relaxation of the assumption of time-invariant noise processes, which is recognized as nontrivial and beyond the scope of the design used by Li, Nychka, and Ammann.

5.1 Extending BHM Methods to Reconstruct Spatially Explicit Climate Data

Extending BHM methods to reconstruct spatially explicit climate data, for example, regularly gridded data comprising climate fields in latitude-longitude space, is a new frontier for significant further effort, following the pioneering work of Tingley and Huybers (2010a, 2010b). As mentioned, a key success of paleoclimatology has been the development of simultaneous whole-field reconstruction approaches, one of the challenges of which is to properly characterize uncertainty. This is a challenge to which application of the BHM approach could be expected to add significant capacity, again both methodologically and as a reference point for improving traditional reconstruction methods. An intriguing possibility in this regard would be to include spatial climate information produced by AOGCMs (and their successors, earth system models, or ESMs) at the process model stage in the BHM, in an analogous manner to the way forcing information is used by Li, Nychka, and Ammann. Another issue that arises for spatial climate reconstructions is that logically they should be weighted in reconstruction skill analyses to take into account the fact the climate system has "nodes" of particular sensitivity or importance for the entire earth (e.g., the tropical Pacific involved in the El Nino-Southern Oscillation phenomenon and the North Atlantic region). Accounting for spatially varying importance of reconstruction success in building validation schemes for spatial climate reconstructions is a nontrivial issue, as it is in weather forecasting, and the application of BHM methods to help in this work could be highly valuable.

5.2 Refining the Data Models Used in BHMs

Finally, significant work could be done to refine the data models used in BHMs for climate reconstruction. The borehole model used by Li, Nychka, and Ammann is realistic at a first-order level, but the tree ring and pollen models are less realistic and deserve significant refinement. Work is currently being done in developing biophysical models of seasonal tree growth driven by a small set of key climatic variables, which carries potential for being incorporated into BHM-based reconstructions (e.g., Evans et al. 2006). The authors are developing a more realistic model for pollen response to temperature, which is described in general terms in Section 5.3 and applied and tested in the Supplemental Online Material (SOM) (cf. Korhola et al. 2002; Haslett et al. 2006). Adding these, and/or other, data model refinements will amplify the inherent strength of the BHM approach in terms of explicit specification of the hierarchical reconstruction model. Of particular interest would be to combine such data model refinements with the incorporation of ESM output at the process model stage, described above. Pursuit of these goals provides important direction for further development work of the kind set forth by Li, Nychka, and Ammann.

5.3 A Binomial-Logistic Pollen Data Model for Use in BHM Reconstruction of Temperature

Modern paleoenvironmental reconstruction from fossil pollen data often attempts to take advantage of the fact that pollen "assemblages" archived in sedimentary deposits are generally rich in taxonomic (i.e., plant type) diversity. To produce a more realistic pollen data model for use in real-world applications of the BHM, however, it was decided to define a "reduced-space" taxonomic model that would be parsimonious in introducing additional parameters that need to be estimated within the BHM. To be even more parsimonious, another possibility is to estimate the pollen model separately, which could then be incorporated in the BHM to serve a role analogous to the transformation matrices utilized by Li, Nychka, and Ammann to define the "forward" (i.e., causal) relationships between climate and proxy information [the "M" matrices in their Equations (4.1)–(4.3)].

To meet these criteria, a refinement of the traditional pollen ratio method (cf. Adam and West 1983) was developed. The ratio method has been known for some time to be a useful tool in pollen-based paleoclimate reconstruction: in situations in which one (or a few) dominant pollen type(s) in a region have a strong positive correlation with a climate variable of interest and another (or a few) dominant pollen type(s) have a strong negative correlation with the same climate variable. A classic example of this situation occurs in the coastal mountain regions of California, where oak (*Quercus*) and pine (*Pinus*) pollen representation, respectively, vary inversely in relation to temperature (Adam and West 1983; Wahl 2003). When counts of these pollen types are combined as $Pinus/(Quercus + Pinus)$ or $Quercus/(Quercus + Pinus)$ ratios, a mathematically appropriate estimation form (which generally has not been utilized in this context) is the binomial logistic generalized linear model (GLM) (Gelman et al. 2004).

The GLM also can readily model this relationship in the forward form of pollen $= g$(climate), which is more physically realistic in terms of the direction of causation. The specification of such a forward model is shown below:

$$r_{\text{num}} \sim \text{Bin}(n, p),$$

where

$$E(r|T) = p = \exp(\eta)/[1 + \exp(\eta)],$$

and

$$\eta = \alpha + \beta(T).$$

Here, r is the pollen ratio formed as above, r_{num} is the ratio numerator, n is the ratio denominator (i.e., the sum of pollen counts), the denominator-specific count is $(n - r_{\text{num}})$, and T is the temperature at each site corresponding to a specific value of r.

As described further in the SOM, α and β were estimated using the GLM algorithm (in the R language), yielding fitted values of $p = E(r|T)$ for given values of T. These fitted $E(r|T)$, T combinations were then compared with the actual sampled r, T pairs to determine how much of the total GLM deviance the $E(r|T)$ values explain, analogous to explained variation in a standard linear model. Importantly, the new ratio method performs extremely well, thereby allowing great simplification from ecological, mathematical, and by extension, computational standpoints. Although it includes only two to four pollen types, it can provide as much or more explained variation in the pollen-temperature relationship as a 64-type "modern analog technique," or MAT (∼80% explained variation in temperate northeast North America where it has been applied; cf. the SOM for description of the MAT). Thus, the new pollen ratio method represents an information-rich, taxonomic "reduced space" data model that can be fruitfully, and efficiently, employed in a BHM framework.

SUPPLEMENTAL MATERIALS

A Pollen 'Forward' Model to Enhance the Realism of the BHM: Provided additional material describing the application and testing of the binomial-logistic pollen data model outlined in Section 5.3. Section S.1: Specification, testing, and estimation of the binomial-logistic GLM. Section S.2: Use of the model in paleotemperature reconstruction and uncertainty estimation. Section S.3: Technical note (regarding languages used in computation). Also provided Figures S1, S2, S3. (supplement.pdf)

ADDITIONAL REFERENCES

Adam, D. P., and West, G. J. (1983), "Temperature and Precipitation Estimates Through the Last Glacial Cycle From Clear Lake, California, Pollen Data," *Science*, 219, 168–170. [904]

Ammann, C. M., Genton, M. G., and Li, B. (2010), "Technical Note: Correcting for Signal Attenuation From Noisy Proxy Data in Climate Reconstructions," *Climate of the Past*, 6, 273–279. [901]

Bürger, G., Fast, I., and Cubasch, U. (2006), "Climate Reconstruction by Regression—32 Variations on a Theme," *Tellus, Ser. A*, 58 (2), 227–235. [902]

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. [901]

Cook, E. R., Woodhouse, C. A., Eakin, C. M., Meko, D. M., and Stahle, D. W. (2004), "Long-Term Aridity Changes in the Western United States," *Science*, 306, 1015–1018. [901]

Evans, M. N., Reichert, B. K., Kaplan, A., Anchukaitis, K. J., Vaganov, E. A., Hughes, M. K., and Cane, M. A. (2006), "A Forward Modeling Approach to Paleoclimatic Interpretation of Tree-Ring Data," *Journal of Geophysical Research*, 111, G03008, DOI: *10.1029/2006JG000166*. [903]

Frank, D., Esper, J., Raible, C. C., Büntgen, U., Trouet, V., Stocker, B., and Joos, F. (2010), "Ensemble Reconstruction Constraints on the Global Carbon Cycle Sensitivity to Climate," *Nature*, 463, 527–530. [901]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC Press. [904]

Hegerl, G. C., Crowley, T. J., Allen, M. R., Hyde, W. T., Pollack, H. N., Smerdon, J. E., and Zorita, E. (2007), "Detection of Human Influence on a New, Validated 1500 Year Temperature Reconstruction," *Journal of Climate*, 20, 650–666. [901]

Imbrie, J., and Imbrie, K. P. (1979), *Ice Ages: Solving the Mystery*, Cambridge, MA and London: Harvard University Press. [901]

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., and Kull, C. (2009), "High-Resolution Palaeoclimatology of the Last Millennium: A Review of Current Status and Future Prospects," *Holocene*, 19 (1), 3–49. [902]

Korhola, A., Vasko, K., Toivonen, H. T. T., and Olander, H. (2002), "Holocene Temperature Changes in Northern Fennoscandia Reconstructed From Chironomids Using Bayesian Modelling," *Quaternary Science Reviews*, 21, 1841–1860. [903]

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. [901]

Mann, M. E., Rutherford, S., Wahl, E., and Amman, C. (2007), "Robustness of Proxy-Based Climate Field Reconstruction Methods," *Journal of Geophysical Research*, 112, D12109, DOI: *10.1029/2006JD008272*; corrigenda (2008), 113, D18107, DOI: *10.1029/2008JD009964*. [901,902]

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F. (2009), "Global Signatures and Dynamical Origins of the Little Ice Age and Medieval Climate Anomaly," *Science*, 326, 1256–1260, DOI: *10.1126/science.1177303*. [901]

Riedwyl, N., Küttel, M., Luterbacher, J., and Wanner, H. (2009), "Comparison of Climate Field Reconstruction Techniques: Application to Europe," *Climate Dynamics*, 32 (2–3), 381–395, DOI: *10.1007/s00382-008-0395-5*. [902]

Rutherford, S., Mann, M. E., Ammann, C. M., and Wahl, E. R. (2010), Comments on "A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness," by B. Christiansen, T. Schmith, and P. Theijll, *Journal of Climate*, 23, 2832–2838, DOI: *10.1175/2009JCLI3146.1*. [902]

Rutherford, S., Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R., Hughes, M. K., and Jones, P. D. (2005), "Proxy-Based Northern Hemisphere Surface Temperature Reconstructions: Sensitivity to Methodology, Predictor Network, Target Season and Target Domain," *Journal of Climate*, 18, 2308–2329. [902]

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*, in press. [902]

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," *Journal of Climate*, 23, 2759–2781. [903]

——— (2010b), "A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison With the Regularized Expectation–Maximization Algorithm," *Journal of Climate*, 23, 2782–2800. [903]

von Storch, H., Zorita, E., and González-Rouco, F. (2009), "Assessment of Three Temperature Reconstruction Methods in the Virtual Reality of a Climate Simulation," *International Journal of Earth Sciences*, 98, 67–82. [902]

Wahl, E. R. (2003), "Pollen Surface Samples for Paleoenvironmental Reconstruction From the Coast and Transverse Ranges of Southern California," *Madroño*, 50 (4), 286–299. [904]

Wahl, E. R., and Ammann, C. M. (2007), "Robustness of the Mann, Bradley, Hughes Reconstruction of Surface Temperatures: Examination of Criticisms Based on the Nature and Processing of Proxy Climate Evidence," *Climatic Change*, 85, 33–69. [901]

Comment

Richard L. SMITH

The paper by Li, Nychka, and Ammann (2010) has exemplified the power of Bayesian Hierarchical Models to solve fundamental problems in paleoclimatology. However, much can also be learned by more elementary statistical methods. In this discussion, we use principal components analysis, regression, and time series analysis, to reconstruct the temperature signal since 1400 based on tree rings data. Although the "hockey stick" shape is less clear cut than in the original analysis of Mann, Bradley, and Hughes (1998, 1999), there is still substantial evidence that recent decades are among the warmest of the past 600 years.

The problem of paleoclimate reconstruction is a natural one for the use of Bayesian hierarchical models (BHMs). As in most BHMs, there is an unobserved "process" which is the true object of interest—in this case, the true series of temperatures. There are also various sources of "data" which are dependent on the "process" with different levels of accuracy—observational data, tree rings, boreholes, ice cores, etc. The problem of paleoclimate reconstruction may be characterized as how to combine the different data series to obtain the best reconstruction of the unobserved process, with suitable measures of uncertainty. The BHM technique is especially valuable for answering nonstandard uncertainty questions, for instance, "what is the probability that the 1990s were the warmest decade of the [1000–2000] millennium?"

In an earlier paper, Li, Nychka, and Ammann (henceforth LNA 2007) used an ensemble reconstruction, obtained via a combination of linear regression, bootstrapping and cross-validation, to reconstruct Northern Hemisphere average temperatures back to 1000, using 14 proxy series first discussed in Mann, Bradley, and Hughes (MBH 1999). Their results showed that there is indeed a high probability that the 1990s were the warmest decade of the millennium. The BHM technique has since been taken up by other authors, such as Tingley and Huybers (2010a, 2010b), Brynjarsdóttir and Berliner (2010), and promises to be the method of choice for future statistical analyses of paleoclimatic data.

In the paper under discussion, LNA (2010) have shown that it is also possible to answer "design"-type questions using BHMs. I believe that this is the logical next step in the scientific application of BHMs to paleoclimatology, and the methodology they have presented will play an important role in the selection of proxies for future paleoclimatological studies. I commend their contribution.

Although I fully support the further development of the BHM approach, it seems to me there is still some merit in looking for simpler statistical approaches, using methods that are routinely taught in first-year graduate courses in statistics, and that can (through the ready availability of the R programming language) be easily adopted by paleoclimatologists without extensive statistical training. Indeed, much of the debate over the "hockey stick curve" has focused on the correct use of elementary statistical methods, in particular, the method of principal components (PCs). For the remainder of this note, I aim to show how routine application of PCs, regression, and time series analysis can be used to resolve some issues that have caused much contention in the literature.

BRIEF SUMMARY OF THE CONTROVERSY

The hockey stick curve, in the form that is currently debated, was first constructed in two papers of MBH (1998, 1999). After

Richard L. Smith is Director, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709-4006, and Mark L. Reed III Distinguished Professor, Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260 (E-mail: *rls@email.unc.edu*). SAMSI is supported by the National Science Foundation, grant DMS 0635449. I am grateful to Doug Nychka and Caspar Ammann for making their data and programs available.

© 2010 American Statistical Association
Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap10507

compiling a large number of proxies, they computed what they claimed was the first PC of the proxy series, with uncertainty bands for the reconstruction of each year's temperature mean.

Their analysis was criticized by a number of authors, but most sharply by McIntyre and McKitrick (2003, 2005a, 2005b) (henceforth, M&M). In particular, M&M (2005a) showed that the version of PC analysis used by MBH was more or less the following:

1. A data source was compiled that, for this discussion, is taken as 70 tree-ring series for 1400–1980.
2. Each series was rescaled so that the mean was 0 and the variance 1 over 1902–1980—the "calibration period" for which real observational data was used.
3. For each series, a linear trend was fitted to 1902–1980, and the series rescaled again by dividing by the standard deviation of the residuals.
4. Based on the rescaled data matrix X, the singular value decomposition $X = UDV^T$ was formed as in conventional PC analysis (however, the X matrix has not been centered over all the series, only over 1902–1980).
5. Based on the SVD, the first PC (PC1) was calculated.
6. Finally, PC1 was rescaled so that the mean was 0 and the variance 1 over 1902–1980. This was the displayed result.

For the current discussion, I have reproduced these results using data and R code provided by Doug Nychka and Caspar Ammann (*http://www.image.ucar.edu/~nychka/Temp/TreePC/*). The basic dataset consists of reconstructed temperatures from 70 trees for 1400–1980, in the North American International Tree Ring Data Base (ITRDB). In Figure 1, the 70 tree ring series have been plotted, after smoothing by applying a moving average filter with weights $w_i = (13 - |i|)/169$, $i = -12, -11, \ldots, 12$, that is, a triangular window over the total span of 25 years with weights summing to 1. This is intended as a smooth representation of the underlying trend and will be used for trend comparisons in the subsequent analysis. As can be seen, there is little visual evidence of an overall upward or downward trend. Figure 2 shows the MBH reconstruction of the individual-year temperatures, together with a smoothed trend computed by the same triangular window. This shows the characteristic hockey stick shape, including a steady rise in temperatures since 1850.

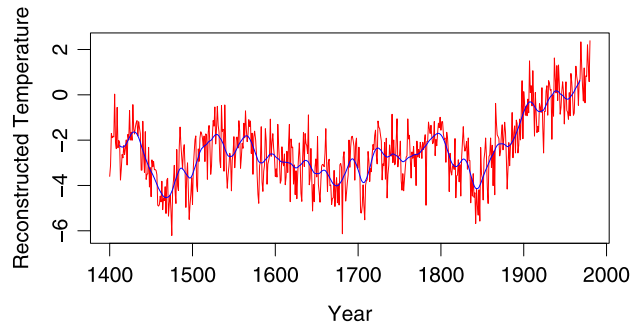

Figure 2. The original temperature reconstruction of MBH (1999), replotted for the present discussion, together with a smoothed trend computed using the same triangular window. The online version of this figure is in color.

Steps 2 and 3 of the above algorithm are nonstandard, since conventional PC analysis centers the data over the whole time period, not just part of it. In additional, conventional PC analysis sometimes standardizes the individual series so that the standard deviation is 1—the standardized and unstandardized analyses are also commonly known as correlation-based and covariance-based PC analysis. M&M (2005a) argued that the version of PC analysis used by MBH can induce a spurious "hockey stick" shape, a point they illustrated by using simulations of red-noise time series with no trend to represent the tree-ring series. In many cases, the simulated series combined with the MBH algorithm resulted in a spurious hockey stick-like curve for the trend. Instead, M&M argued for a conventional PC analysis with centering over the whole time series. As illustration, Figure 3 shows PC1 from a correlation-based PC analysis, with the smooth trend using the same moving average smoother as previously. The output series has been rescaled to have mean 0 and variance 1 over the 1902–1980 period, to allow direct comparison with MBH. As noted by M&M, this series does not support the notion that recent decades were substantially warmer than temperatures in earlier centuries.

In a report to the House of Representatives of the U.S. Congress, Wegman, Scott, and Said (2006) confirmed the M&M results and gave much additional discussion, generally supporting the conclusion of M&M that there was no hockey stick shape. At the same time, the National Research Council

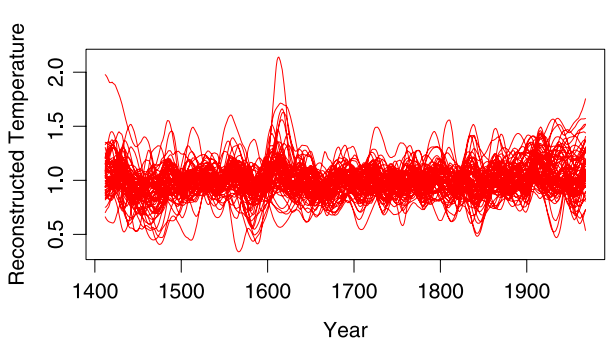

Figure 1. Reconstructed temperatures from 70 tree rings (1400–1980) in the North American ITRDB dataset. Each series has been smoothed using the 25-year triangular window described in the text. The online version of this figure is in color.

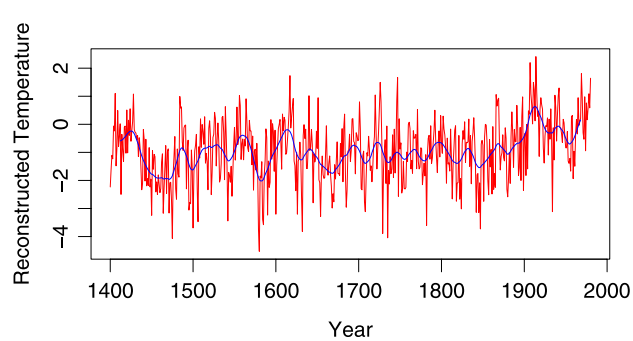

Figure 3. The first principal component, computed from the tree ring dataset using a conventional (correlation-based) PC decomposition, together with the smoothed trend. The online version of this figure is in color.

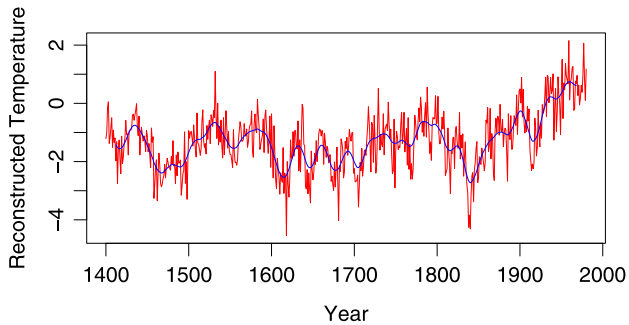

Figure 4. The second principal component, computed as in Figure 3. The online version of this figure is in color.

commissioned a report, published as NRC (2006). They acknowledged the high level of uncertainty in reconstructions of medieval temperatures, but concluded that the balance of evidence, from many different studies, still supported the overall hockey stick shape of the temperature curve.

One criticism of the M&M analysis is their focus on PC1 as the main indicator of a signal in the series. M&M (2005b, pp. 75–76) acknowledged this point, stating that a hockey stick shape reappears in the second PC of a standardized analysis (see Figure 4, where the 25-year smoother is also included), and the fourth PC of an unstandardized analysis, but argued against this interpretation, saying (with respect to the unstandardized analysis) that MBH's "conclusion about the uniqueness of the late 20th century climate hinges on the inclusion of a low-order PC series that only accounts for 8% of the variance of one proxy roster." In contrast, Ammann and Wahl (2007), Wahl and Ammann (2007) argued for the "robustness" of the MBH results, in particular the "convergence" of the reconstructed curves as the number of PCs in the reconstruction increased, and that this convergence holds regardless of whether the PCs are constructed using standardized series, unstandardized series, or the MBH technique. However, they did not discuss the standard error of the reconstructed curve, nor provide a formal criterion for selecting the number of PCs to include in the reconstruction.

PRINCIPAL COMPONENTS REGRESSION

Suppose we have observed temperatures y_t for $t = 1902, \ldots, 1980$, and proxy series $\{x_{jt},\ j = 1, \ldots, q\}$ (where, here, $q = 70$) for $t = 1400, \ldots, 1980$. A natural way to think about the paleoclimate reconstruction problem is first to fit the regression

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j x_{jt} + \epsilon_t \tag{1}$$

for 1902–1980, and then, having estimated coefficients $\hat{\alpha}_j,\ j = 0, \ldots, q$, to apply the fitted regression curve, $\hat{y}_t = \hat{\alpha}_0 + \sum_{j=1}^{q} \hat{\alpha}_j x_{jt}$, to reconstruct the temperature curve prior to 1902.

Direct application of (1), however, is open to the objection that the number of regressors (70) is very close to the number of data points used to fit the regression (79), creating potentially serious overfitting problems. A standard method for dealing with this problem is first to transform the covariates $\{x_{jt},\ j = 1, \ldots, 70\}$ to PCs $\{u_{kt},\ k = 1, \ldots, 70\}$, ordered by decreasing variance. Then, we fit the observed temperatures to a subset of the PCs,

$$y_t = \beta_0 + \sum_{k=1}^{K} \beta_k u_{kt} + \epsilon_t, \tag{2}$$

where K is to be determined. The argument for this is that with moderate K, chosen to capture most of the variability in the covariates, the right-hand side of (2) contains almost as much information as the right hand side of (1), but with far fewer regression coefficients to be estimated. For the moment, I assume a conventional ordinary least squares (OLS) regression analysis in which the ϵ_t are uncorrelated with mean 0 and common unknown variance σ^2.

This method was applied with the observed series y_t taken as the "HADCRUT3" global temperature mean anomalies of the Climate Research Unit of the University of East Anglia (*http://www.cru.uea.ac.uk/cru/data/temperature*). Based on the PC regression, I compute the predicted temperature series $\hat{y}_t = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k u_{kt}$, and also its smoothed version $\tilde{y}_t = \sum_{i=-12}^{12} w_i \hat{y}_{t-i}$. Noting that we can also write $\tilde{y}_t = \sum_{k=0}^{K} \hat{\beta}_k \tilde{u}_{kt}$ where $\tilde{u}_{kt}$ is 1 for $k = 0$, $\sum_{i=-12}^{12} w_i u_{k,t-i}$ for $k = 1, \ldots, K$, the prediction intervals for $\tilde{y}_t$ can be computed by the standard formula given in regression textbooks. Therefore, we can display the smoothed reconstructed curve $\tilde{y}_t$, together with $100(1-\alpha)\%$ prediction bounds for any α, pointwise for each t. In the following examples, I have taken $\alpha = 0.1$ to give 90% prediction intervals.

Figure 5 shows the resulting reconstruction for $K = 1, 2, 5, 10, 20, 50$. The curve for $K = 1$ is flat, suggesting that the first PC has very little predictive power. However, each of the remaining curves has a noticeable hockey stick shape. For $K = 50$, and to a lesser extent $K = 20$, the prediction intervals are extremely narrow in the portion of the series with observational data (1902–1980), a sure sign of overfitting. For $K = 2, 5, 10$, however, the curves look very similar and there is no clear-cut choice among them.

A more systematic comparison may be made by computing various measures of fit. In comparing regression and/or time series models, three common criteria for selecting the best model are the Akaike Information Criterion (AIC; Akaike 1973), the Bayesian information criterion (BIC; Akaike 1978) and the bias-corrected Akaike Information Criterion (AICC; Hurvich and Tsai 1989). For the PC regression model with K up to 10, but still for the moment without making any allowance for time series autocorrelations, these criteria are listed in Table 1.

It can be seen that as K increases, each of AIC, BIC, AICC drops sharply (indicating improved fit) at $K = 2$ and again at $K = 8$. The minimum is at $K = 9$ for AIC and AICC and $K = 2$ for BIC: as typically happens, BIC chooses a more parsimonious model.

The omission of autocorrelation from the foregoing analyses is potentially a serious matter, since the width of the prediction interval could be substantially larger if autocorrelation is included. Therefore, I now extend the previous analysis to include a time series component.

The logical extension to the assumption that the errors ϵ_t in (1) or (2) are independent is that they form an ARMA(p, q) time series for suitable p and q (see, e.g., Brockwell and Davis 2003 for extensive discussion of ARMA modeling). The most

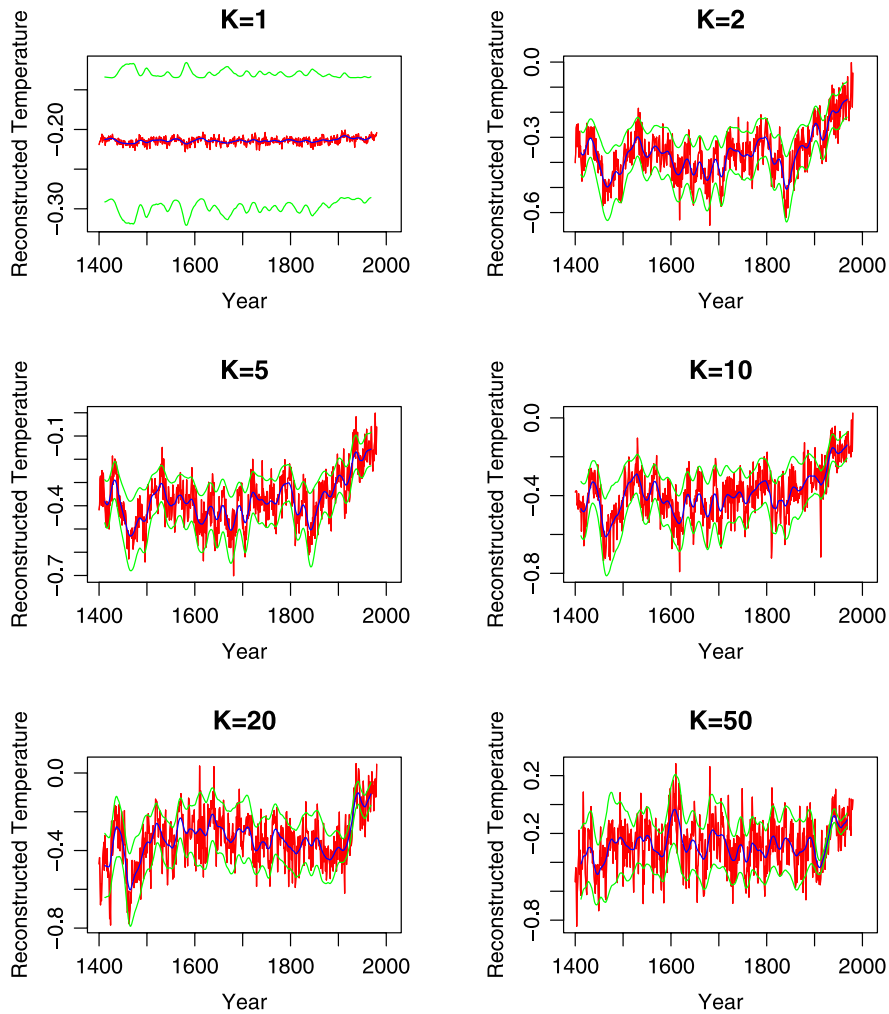

Figure 5. Six reconstructions of historical temperature anomalies, together with their smoothed trends and pointwise 90% prediction intervals on the trends. The online version of this figure is in color.

powerful way to select a model is to treat all of K, p, and q as undetermined model parameters, to perform a generalized least squares (GLS) analysis, and to select K, p, and q to minimize one of AIC, BIC, or AICC. Unfortunately, this procedure quickly produces unwieldy models and does not lead to a clear-cut conclusion. For example, fitting models by minimizing AIC up to $K = 9$, $p = 10$, $q = 5$ produced the best model at $K = 8$, $p = 2$, $q = 5$. These results were obtained by maximum likelihood fitting using the `arima` command in R (R Core Development Team 2010), ignoring models for which the MLE algorithm did not converge or for which the resulting model fit violated the stationarity condition for the autoregressive part of the model. However, the final model is hard to interpret with so many parameters, and it seems probable that still higher-order models would be obtained if larger values of K, p, q were permitted. Similar results were obtained using BIC and AICC.

Table 1. Table of AIC, BIC, AICC values for OLS regression without allowing for autocorrelation

K	AIC	BIC	AICC
1	−40.4	−33.3	−40.1
2	−58.2	−48.7	−57.7
3	−58.7	−46.9	−57.9
4	−57.3	−43.0	−56.1
5	−55.4	−38.8	−53.8
6	−57.3	−38.4	−55.3
7	−58.1	−36.8	−55.5
8	−63.8	−40.1	−60.5
9	−66.4	−40.4	−62.5
10	−66.1	−37.7	−61.4

As an alternative to full GLS time series regression, therefore, I used the same OLS fits for the regression components produced earlier, but selected the optimal ARMA(p, q) model fitted to the residuals, and then recalculated the width of the prediction intervals to take account of the autocorrelation. This produced more easily interpretable results. For example, with $K = 2$, the optimal ARMA model had $p = 1$, $q = 2$ when selecting by AIC and $p = 1$, $q = 0$ when using BIC or AICC. With $K = 9$, all three selection criteria resulted in AR(1) as the optimal time series model—incidentally relevant to LNA (2010), where they used AR(2) as the time series model for residuals, though Dr. Li remarked in her oral presentation that the AR(1) model appears equally suitable in practice.

For the three models just derived, the reconstructed curves, with prediction interval bounds for the 25-year moving average,

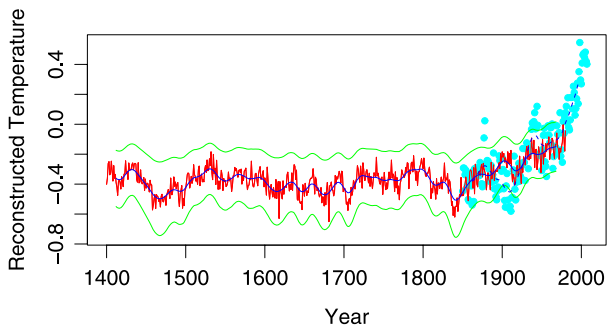

Figure 6. Reconstructed series with time series corrections: $K = 2$ PCs with AR(1) residuals. The solid curves represent the smoothed reconstructed series with pointwise 90% prediction intervals. The dashed curve at the right-hand end is the same smoother applied to the observational data points. The online version of this figure is in color.

are shown in Figures 6 through 8. Also shown on the plot are the actual global mean temperatures for 1902–1980, and a 25-year triangular moving average filter applied to those. The three figures look very similar to each other, though the width of the prediction intervals is about twice that in Figure 5. All three confirm that the reconstructed smoothed temperature for prior centuries was well below its value in recent decades.

DISCUSSION AND SUMMARY

This analysis has used principal components regression combined with time series analysis of the residuals to reconstruct the global mean temperature series back to 1400. I smoothed the reconstructed series using a 25-year triangular moving average, and calculated 90% prediction intervals on the smoothed reconstruction as a measure of uncertainty. Three standard statistical model selection criteria (AIC, BIC, and AICC) were used to select the model orders K (number of PCs), p and q (for the autoregressive and moving average components of the time series model fitted to the residuals). Although these criteria do not lead to clear-cut selection of the best model, the final reconstructions do not appear to depend too sensitively on the model selected. Taking into account the general desire in applied statistics for a parsimonious model, the model with $K = 2$ PCs and AR(1) residuals appears adequate.

The idea of PC regression as a technique in paleoclimate reconstruction is not new—for example, it was discussed in chapter 9 of NRC (2006)—but it does not appear to have been systematically developed.

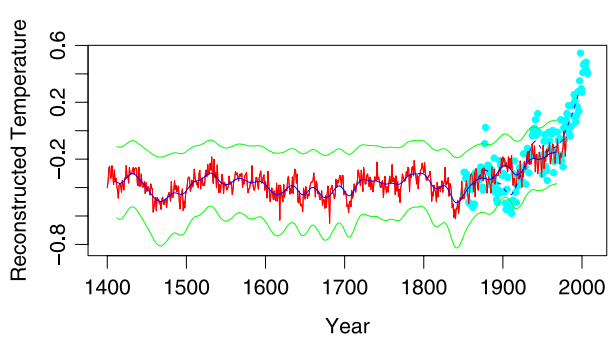

Figure 7. Reconstructed series with time series corrections: $K = 2$ PCs with ARMA(1, 2) residuals. The online version of this figure is in color.

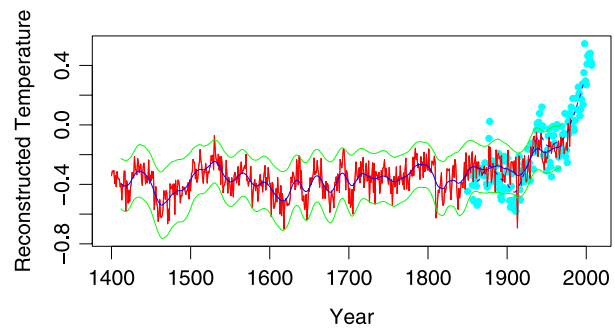

Figure 8. Reconstructed series with time series corrections: $K = 9$ PCs with AR(1) residuals. The online version of this figure is in color.

The results support an overall conclusion that the temperatures in recent decades have been higher than at any previous time since 1400. On the other hand, none of the recent reconstructions shows as sharp a hockey stick shape as the widely reproduced figure 3(a) of MBH (1999), so in that respect, critics of the hockey stick are also partially vindicated by these results.

I have confined this discussion to statistical aspects of the reconstruction, not touching on the question of selecting trees for the proxy series (extensively discussed by M&M, Wegman, Scott, and Said and Ammann/Wahl) nor the apparent recent "divergence" of the relationship between tree ring reconstructions and measured temperatures (see, e.g., NRC 2006, pp. 48–52). I regard these as part of the wider scientific debate about dendroclimatology but not strictly part of the statistical discussion, though it would be possible to apply the same methods as have been given here to examine the sensitivity of the analysis to different constructions of the proxy series or to different specifications of the starting and ending points of the analysis.

ADDITIONAL REFERENCES

Akaike, H. (1973), "Information Theory and an Extention of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267–281. [907]

——— (1978), "A Bayesian Analysis of the Minimum AIC Procedure," *Annals of the Institute of Statistical Mathematics*, 30, 9–14. [907]

Brockwell, P. J., and Davis, R. A. (2003), *Introduction to Time Series and Forecasting* (2nd ed.), New York: Springer. [907]

Brynjarsdóttir, J., and Berliner, L. M. (2010), "Beysian Hierarchical Modeling for Paleoclimate Reconstruction From Geothermal Data," preprint, Ohio State University. [905]

Hurvich, C. M., and Tsai, C.-L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307. [907]

Li, B., Nychka, D. W., and Ammann, C. M. (2010), "The Value of Multi-Proxy Reconstruction of Past Climate," *Journal of the American Statistical Association*, 105, 883–895. [905,908]

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. [905,906,909]

McIntyre, S., and McKitrick, R. (2003), "Corrections to the Mann et al. (1998) Proxy Data Base and Northern Hemisphere Average Temperature Series," *Energy and Environment*, 14, 751–771. [906]

——— (2005a), "Hockey Sticks, Principal Components, and Spurious Significance," *Geophysical Research Letters*, 32, L03710, DOI: *10.1029/2004GL021750*. [906]

——— (2005b), "The M&M Critique of the MBH98 Northern Hemisphere Climate Index: Update and Implications," *Energy and Environment*, 16, 69–100. [906,907]

R Core Development Team (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, available at *http://www.R-project.org*. [908]

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," *Journal of Climate*, 23, 2759–2781. [905]

——— (2010b), "A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison With the Regularized Expectation–Maximization Algorithm," *Journal of Climate*, 23, 2782–2800. [905]

Wahl, E. R., and Ammann, C. M. (2007), "Robustness of the Mann, Bradley, Hughes Reconstruction of Northern Hemisphere Surface Temperatures: Examination of Criticisms Based on the Nature and Processing of Proxy Climate Evidence," *Climatic Change*, 85, 33–69. [907]

Wegman, E. J., Scott, D. W., and Said, Y. H. (2006), "Ad hoc Committee Report on the 'Hockey Stick' Global Climate Reconstruction," presented to the Committee on Energy and Commerce and the Subcommittee on Oversight and Investigations, U.S. House of Representatives. [906]

Rejoinder

Bo LI, Douglas W. NYCHKA, and Caspar M. AMMANN

Paleoclimate reconstruction provides a very good example for the necessity of combining statistics with the geosciences and obviously both areas benefit from this combination. The "Climategate" review panel chaired by Lord Oxburgh concluded that "*It would be helpful for researchers to work more closely with professional statisticians in future. This would ensure the best methods were used when analyzing the complex and often 'messy' data on climate*," quoted from BBC news (*http://news.bbc.co.uk/2/hi/science/nature/8618024.stm*). In our paper we provided a framework for the paleoclimate reconstruction to help more statisticians contribute to this area and also to provide a clearer description of the statistical assumptions that support a particular method. In this way we hope that we have addressed needs highlighted by this recent review panel. The key points from the three discussions are the point out limitations of our analysis and also solutions for more improving the reconstructions. We note that there is a consensus among the discussants that the Bayesian hierarchical model (BHM) approach is a positive contribution to paleoclimate and the numerous constructive comments are much appreciated. We also thank the discussants for sharing some new work in their discussions on the spatial aspects of reconstructions and on better models for linking proxies to climate variables. To keep our response succinct we will not comment on this new work except to acknowledge the value and innovativeness of these extensions.

1. DESIGN OF THE EXPERIMENT

Cressie and Tingley (CT) examined our experiment and interpreted our set up as having only one treatment—the BHM applied to the dataset. Indeed, we do not compare this with other methods of reconstruction. Our main interest was to understand the information content of different kinds of climate proxies and how this interacted with a process model. For that purpose, we consider the combination of three factors of with/without forcing covariates, with/without proxy noise and modeling $\mathbf{T}_1/\mathbf{T}$. In terms of experimental design language we have five blocks comprising the different kinds of models and a 2^3 design for the treatments. We accept CT's criticism on the lack of replication. To our knowledge, however, this climate model run is unique in both its resolution and length. Adding in another climate model simulation would be better described as creating another factor for the experimental design rather than a replicate. Simulating a sample (also termed an ensemble in the geosciences) of climate model runs that reflect the model uncertainty and the variation in response is a grand challenge not only for paleoclimate studies but also for projections of future climate change.

CT questioned why the "oracle" case is not always the best? If forcings are not included the "oracle" is unanimously the best in both measures. When forcings are included, the "oracle" is sometimes slightly worse than the others. The most likely explanation for this result is the sampling error due to considering just one replication and this reinforces the point that we should consider more than one climate model run for evaluating the BHM approach.

2. REFINEMENT OF DATA MODELS

We agree that the forward models are unknown and much more complex than our current models, which simply are linear approximations and are not derived from physics. Regarding the moving average calculated in the forward model for tree rings, CT suggested to use the moving window as $t-10$ to t instead of the current $t-5$ to $t+5$. We agree that for a single tree ring this does not make physical sense but could reflect a composite of different trees or another form of annual climate proxy. The "tree ring" proxy sited at 75°N in our study was actually intended to represent an ice-core based proxy that had the same annual time scale information as a tree-ring proxy.

Both CT and Wahl, Schoelzel, and Tigrek (WST) suggested a seasonal tree-growth model (Evans et al. 2006) as the future forward model for tree rings. This is indeed a much more realistic model to consider, but it requires several other climate inputs including precipitation, transpiration, runoff, and solar radiation, and it will introduce some nonlinear data models. Therefore, we suspect a simpler version of this growth model might be more useful in practice. WST have made impressive progress in developing a binomial-logistic generalized linear model as an efficient forward model for pollen and we look forward to these components being incorporated into spatial reconstructions.

Bo Li is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906 (E-mail: *boli@stat.purdue.edu*). Douglas W. Nychka is Senior Scientist and Director of Institute for Mathematics Applied to Geosciences (E-mail: *nychka@ucar.edu*) and Caspar M. Ammann is scientist (E-mail: *ammann@ucar.edu*), National Center for Atmospheric Research (NCAR), Boulder, CO 80307.

© 2010 American Statistical Association
Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap10450

CT raised the concern that the white noise was added onto GCM *before* the forward matrices was applied is not consistent with the models for tree rings (4.1) and pollen (4.2). Since the transformer $\mathbf{M}_D$ retains only high-frequency variability, the white noise added to GCM temperature will be carried over to the synthetic tree rings. Therefore, there is little effect to synthetic tree rings whether the white noise is added before or after the transformer is applied, and the AR(2) is used to model a mixture of white and serially correlated noise. The additive error term in (4.2) serves an approximation to the well-known dating error in pollen rather than for the purpose of modeling the white noise.

As opposed to assuming known forward models, CT suggested an approach to parametrize the unknown transformation matrices using M_D for example. They proposed to have $D_i(t) = \mu_i + c_1 T(t-5) + \cdots + c_{11} T(t+5)$ and then estimate the parameters with data. However, in order to keep the parsimony, that model imposes a strong restriction that the relationship between any tree-ring and global temperatures share the same slope parameters, which sounds a bit unrealistic as implied in the "divergence" problem that CT mentioned in Section 4.3 of their discussion. Their another suggestion of estimating transformation matrices by putting a prior on them and then carrying out a sensitivity analysis is more general and an excellent compromise between flexibility and controlling the number of model parameters.

3. SPATIAL RECONSTRUCTION

CT has sketched the hierarchies and presented a space–time random effects model that proposes nonseparable and nonstationary spatio-temporal covariances as expected over a global spatial scale and a millennial temporal scale (Cressie, Shi, and Kang 2010). This model also allows for fast computation of iterations in MCMC by reducing the dimension of basis functions, and is a promising, feasible approach to deal with data over a large spatial scale. Besides, WST raised excellent points that spatial climate information produce by AOGCMs can be included in the process stage, and the spatial reconstruction should be weighted to account for "modes" in the Earth system with particular sensitivity for the entire earth. The spatial network sensitivity and spatial network design for proxies mentioned in CT are definitely interesting research questions that we should consider. In particular, what are the optimal locations to look for new proxies? Or, given all the available data, which locations should be considered?

4. OTHER ISSUES WITH BHM MODELS

Responding to CT's comments concerning the forcings, we only introduced observational errors into the volcanic forcing because this reflected our understanding that errors in the other two forcings have much less impact. The discrepancy in different estimates of solar irradiance $\mathbf{S}$ only amounts to a different scaling, which will be absorbed into the parameter β_2. The greenhouse gas estimates are nearly flat prior to 1950s and then afterwards are the direct measurements, therefore the uncertainty in greenhouse gases are expected to be small and thus negligible. CT suggested the modeling of measurement errors in instrumental temperatures, motivated by the various noises added to instrumental temperatures in the post processing (Brohan et al. 2006). We find this a perceptive comment. In this sense the instrumental record is just another climate proxy, one that we hope has a small measurement error and a simple forward model! Including the instrumental as another observational equation is an elegant way to deal with all the observed information in a unified way.

Due to the fact that the radiative forcing associated with CO_2 increases as the log of the mixing ratio and also the form of multiplicative error in volcanism, CT suggested a new linear process model:

$$\mathbf{T} = \beta_0 \mathbf{1} + \beta_1 \mathbf{S} + \beta_2 \log(\mathbf{V}_0) + \log(\mathbf{C}) + \epsilon.$$

CT also suggested to investigate the role of each forcings as our systematic evaluation for each proxy. Both are valuable suggestions that are worthwhile to try out in future studies.

Considering the relatively extensive statistical training required to use the BHM and for convenience of paleoclimatologists, Smith discussed the efficiency and benefits of using principal components regression combined with time series analysis. We agree that with simple data structure, regression techniques can be very useful and much appreciated for their merits of being easy to understand and implement, but it is also worth mentioning that those techniques may suffer from the attenuation effects caused by errors-in-variables (Ammann, Genton and Li 2010). A careful examination of assumptions made for regressions and appropriate corrections can perhaps improve those methods.

In summary, the discussions have sketched some important improvements to our work and future directions. We are glad our paper has initiated these valuable responses.

ADDITIONAL REFERENCES

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, DOI: *10.1029/2005JD006548*. [911]

Cressie, N., Shi, T., and Kang, E. L. (2010), "Fixed Rank Filtering for Spatio-Temporal Data," *Journal of Computational and Graphical Statistics*, 19, 724–745. [911]

Evans, M., Reichert, B., Kaplan, A., Anchukaitis, K., Vaganov, E., Hughes, M., and Cane, M. (2006), "A Forward Modeling Approach to Paleoclimatic Interpretation of Tree-Ring Data," *Journal of Geophysical Research*, 111, DOI: *10.1029/2006JG000166*. [910]