Comment: Hierarchical Statistical Modeling for Paleoclimate Reconstruction

(Discussion of "The value of multi-proxy reconstruction of past climate"
by Bo Li, Dougles W. Nychka, and Caspar M. Ammann,
to appear in the *Journal of the American Statistical Association)*

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 1,000 Years Before Present) climate from multi-proxy 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 multi-proxy 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

*Noel Cressie (e-mail address: ncressie@stat.osu.edu) is Director of the Program in Spatial Statistics and Environmental Statistics, Professor of Statistics, and Distinguished Professor of Mathematical and Physical Sciences at The Ohio State University, Columbus, OH. Martin P. Tingley (e-mail address: mtingley@samsi.info) is Post-Doctoral Fellow at the Statistical and Applied Mathematical Sciences Institute (SAMSI). 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 No. DSM-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.

through (parametric) probability distributions. This might not be easy for a scientist to do, and hence the statistician's involvement is needed in the paleoclimate investigation 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 (that is, 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.

Posterior distributions from an HM require 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 the introduction.

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? Their 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 et al. [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 Multi-proxy 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.

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 et al., 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 supplement 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 "lab animal" would be closer in make-up to the "patient").

Pollen pseudo-proxies are created by averaging the model output over a number of $7.5^\circ \times 7.5^\circ$

regions, adding noise, removing 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 temporal 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 $20^\circ \times 20^\circ$ 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 500m, and then they sample this depth profile every 5m. 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 $20^\circ \times 20^\circ$ 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 likely motivated by the observation that surface temperatures averaged over longer time scales tend to reflect larger spatial scales.

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. 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 estimates of forcing time series, 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. Let us 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 animals" are different from the "patients."

4 The HM used for reconstructing past climate

4.1 Heirarchical 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 distributions are in fact *conditional* on $\mathbf{S}$ and $\mathbf{C}$. This is an assumption that we invite LNA to comment on. Was the reason 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 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, the notation, $\mathbf{T}_1$, for past temperature is 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 the appropriate goal should be to make inference on the *unknown* $\mathbf{T}_{1,0}$, given the temperature *data* $\mathbf{T}_2$.

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

conditional on $\mathbf{T}_0$, and the focus of the study is on gaining knowledge about the unobserved past climate, $\mathbf{T}_{1,0}$, 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 (Fig. 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 (Fig. 2 of LNA). First, a number of the tree-ring pseudo-proxies are located in the Southern Hemisphere, despite the goal being 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 Eqn. (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 $\mathbf{C}_0$); according to the IPCC's Fourth Assessment report, this assumption is reasonable [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 Fig. 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 Eqn. (4.5) be modified to be linear 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 method. 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.)

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 LNA 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; we return to this point below.

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 (Eq. (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 experiments.

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 $\{\mathbf{T}_{0,1}(\mathbf{s})\colon \mathbf{s} \in \textit{globe}\}$, $\{\mathbf{T}_{0,2}(\mathbf{s})\colon \mathbf{s} \in \textit{globe}\}$, ..., 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.

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 et al. [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

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

where $\{S_t\colon t = 1, 2, \ldots, T\}$ are made up of known spatial basis functions and $\{\boldsymbol{\eta}_t\colon t = 1, 2, \ldots, T\}$ is an r-dimensional *vector autoregressive* time series. Importantly, r is fixed. In Cressie et al. [2010], r was on the order of 100. The last term in the STRE model, $\{\boldsymbol{\xi}_t\colon t = 1, 2, \ldots\}$, 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,b] 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 $2,000$ observatons 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. 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. The suggestions we have made are in support of the HM approach, and LNA's paper shows that an implementation of an HM 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.

8 Acknowledgements

References

W.J. Aldworth and N. Cressie. Sampling designs and prediction methods for Gaussian spatial processes. In S. Ghosh, editor, *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pages 1–54. Marcel Dekker, New York, 1999.

K.R. Briffa, P.D. Jones, T.S. Bartholin, D. Eckstein, F.H. Schweingruber, W. Karlen, P. Zetterberg, and M. Eronen. Fennoscandian summers from AD 500: Temperature changes on short and long timescales. *Climate Dynamics*, 7(3):111–119, 1992.

K.R. Briffa, F.H. Schweingruber, P.D. Jones, T.J. Osborn, S.G. Shiyatov, and E.A. Vaganov. Reduced sensitivity of recent tree-growth to temperature at high northern latitudes. *Nature*, 391 (6668):678–682, 1998.

P. Brohan, J.J. Kennedy, I. Harris, S.F.B. Tett, and P.D. Jones. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, 2:99–113, 2006.

B. Christiansen, T. Schmith, and P. Thejll. A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. *Journal of Climate*, 22(4):951–976, 2009.

N. Cressie, T. Shi, and E.L. Kang. Fixed Rank Filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 2010. Forthcoming.

J. Esper, E.R. Cook, and F.H. Schweingruber. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science*, 295(5563):2250, 2002.

M. Evans, B. Reichert, A. Kaplan, K. Anchukaitis, E. Vaganov, M. Hughes, and M. Cane. A forward modeling approach to paleoclimatic interpretation of tree-ring data. *Journal of Geophysical Research*, 111, 2006.

R.A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK, 1935.

P. Forster, V. Ramaswamy, P. Artaxo, T. Berntsen, R. Betts, D.W. Fahey, J. Haywood, J. Lean, D.C. Lowe, G. Myhre, J. Nganga, R. Prinn, G. Raga, M. Schulz, and R. Van Dorland. Changes in atmospheric constituents and in radiative forcing. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor, and H.L. Miller, editors, *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, Cambridge, UK, 2007.

P.D. Jones, K.R. Briffa, T.J. Osborn, J.M. Lough, T.D. van Ommen, B.M. Vinther, J. Luterbacher, E.R. Wahl, F.W. Zwiers, M.E. Mann, et al. High-resolution palaeoclimatology of the last millennium: A review of current status and future prospects. *The Holocene*, 19(1):3, 2009.

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

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

T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

A.V. Shashkin and E.A. Vaganov. 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(5):275–280, 1993.

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

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