A Bayesian approach to reconstructing climate fields from proxy data

Analysis of a 600 year multi proxy temperature data set

Martin Tingley (tingley@fas.harvard.edu), Peter Huybers (Harvard University), Konrad A Hughen (WHOI)

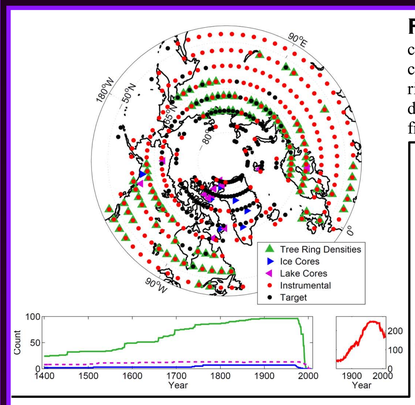

Figure 1: Data. The instrumental data set consists of 249 temperature time series from the CRU compilation. The proxy data set is composed of 96 tree ring density (TRD) series, 13 annual lake floor sediment deposit thickness series (LAKE), and seven δO^{18} series from ice cores from ice cores (ICE).

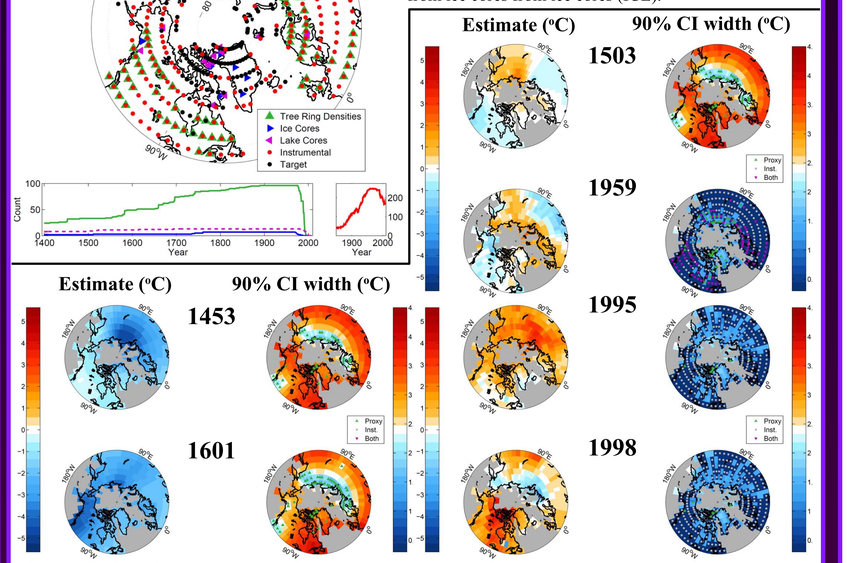

Figure 2 : Fields. Estimates of the field for selected years, along with the widths of the 90% credible intervals. These quantities are calculated at each location as the median and distance between the 5th and 95th percentiles of the posterior draws for that year, respectively. The spatial mean in both 1503 and 1959 is nearly zero, 1453 and 1601 are cold years, and both 1995 and 1998 are warm. In general, the uncertainty grows with distance from the nearest observations, while the field estimates have more spatial structure for years with more observations. These features are a result of the specified spatial covariance form.

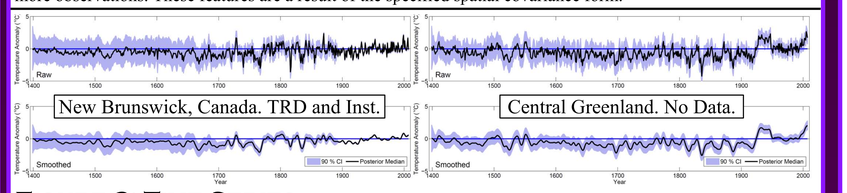

Figure 3: Time Series. Upper panels: black corresponds to the median of the posterior draws, while the shading indicates the 5th and 95th percentiles. In New Brunswick, instrumental observations are available from 1869-1870 and from 1892-2007, and a tree ring density time series covers the interval 1697-1976. Lower panels: estimates of the time series smoothed by a nine point Hanning window, and the uncertainty in that quantity, calculated by applying the smoothing function to the spatial time series resulting from each ensemble member, and then calculating percentiles.

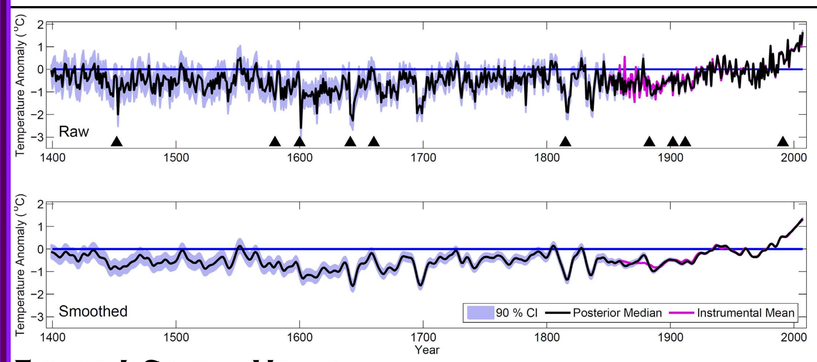

Figure 4: Spatial Means. As in Figure 3, but for the spatial mean. Black triangles in the upper panel indicate volcanic eruptions with a VEI of at least 6.

1 Abstract

We present a Bayesian model to assimilate incomplete (in space and time) instrumental and proxy data sets to estimate, with uncertainties, the time evolution of a climate field. The Bayesian model consists of a process level that describes the evolution of the true climate field as a multivariate AR(1) process with spatially correlated innovations; a data level that specifies the relationships between the measurements (proxies and instrumental) and the true field values; and a prior level that specifies diffuse prior distributions for all unknown parameters. Multiple draws from the posterior produce an ensemble of spatially and temporally complete realizations of the field, compatible with the data and the model assumptions. Probability distributions for various statistics can be estimated from this ensemble, from simple measures like the time series of spatial means to more exotic quantities like the probability that a given year featured the most extreme value of the climate field during the reconstruction.

We demonstrate the utility of this approach via an analysis of a 600 year surface temperature reconstruction for high Northern latitudes based on tree ring, ice core, and lake sediment core data, as well as the Climate Research Unit's [CRU; Brohan et al., 2006] compilation of instrumental observations.

2 Analysis Model

x

For ease of description we will assume the field of interest is that of annual mean surface temperatures, though the method we have developed is general and in principle applicable to the reconstruction of any climate field.

The evolution of the true temperature field sampled at a finite number of spatial locations, $\mathbf{T}_t$, is assumed to follow a multivariate first order auto regressive process:

$$\mathbf{T}_t - \mu\mathbf{1} = \alpha\left(\mathbf{T}_{t-1} - \mu\mathbf{1}\right) + \epsilon_t \tag{1}$$

where μ is the mean of the process, α is the AR(1) coefficient, and the subscript t indexes the year. The innovations, ϵ_t, are assumed to be independent and identically distributed (iid) normal draws, $\epsilon_t \sim N(0, \Sigma)$, with spatial covariance structure given by:

$$\Sigma_{ij} = \sigma^2 \exp\left(-\phi|\mathbf{x}_i - \mathbf{x}_j|\right), \tag{2}$$

where $|\mathbf{x}_i - \mathbf{x}_j|$ is the distance between the ith and j^{th} elements of the field vector $\mathbf{T}$.

It is useful to decompose the vector $\mathbf{T}$, at each year, into three sub-vectors:

$$\mathbf{T} = (\mathbf{T}_I, \mathbf{T}_P, \mathbf{T}_R) \tag{3}$$

where $\mathbf{T}_I$ and $\mathbf{T}_P$ are the true temperatures at locations for which there are instrumental or proxy observations, respectively. If there is both a proxy and instrumental observation at the same location, then the true field value for that location appears in both $\mathbf{T}_I$ and $\mathbf{T}_P$. $\mathbf{T}_R$ are the true temperatures at the target locations where there are no observations. For the analysis presented below, we select these target locations to be the remaining nodes of a regular grid.

The instrumental observations at each year, $\mathbf{W}_{I,t}$, are assumed to be noisy versions of the true temperatures at the corresponding locations:

$$\mathbf{W}_{I,t} = \mathbf{T}_{I,t} + \mathbf{e}_{I,t}. \tag{4}$$

The noise terms are assumed to be iid multivariate normal draws, $\mathbf{e}_{I,t} \sim N\left(0, \tau_I^2\mathbf{I}\right)$, where $\mathbf{I}$ is the identity matrix. Note that the instrumental temperature observations are subject to systemic errors [Brohan et al., 2006] which are not dealt with here.

The proxy observations are assumed to have an unknown, statistically linear relationship to the true temperatures at the corresponding locations, which motivates the regression equation:

$$\mathbf{W}_{P,t} = \beta_1\mathbf{T}_{P,t} + \beta_o\mathbf{1} + \mathbf{e}_{P,t}. \tag{5}$$

The noise terms are once more assumed to be iid normal draws, $\mathbf{e}_{P,t} \sim N\left(0, \tau_P^2\mathbf{I}\right)$. We specify a separate proxy observation equation for each type of proxy data.

In order to close the analysis scheme, priors must be specified for the scalar parameters and the climate field for the first year of the analysis. Our approach is to use weakly informative but proper prior distributions, and show that the information provided by the data overwhelms the prior. A package of Matlab code, which we have dubbed BARSAT for "A Bayesian Algorithm for Reconstructing Spatially Arrayed Temperatures," is then used to infer the joint distribution of the scalar parameters that define the model and the field values through time at the target locations. A more through description of this algorithm can be found in Tingley and Huybers [2009a,b], and the code package is freely available at `people.fas.harvard.edu/~tingley/`.

3 A 600 year temperature data set

We use BARSAT to infer land surface temperatures between 45°N and 85°N over the last 600 years.The basic result of applying BARSAT to a data set is an ensemble of posterior draws of the space-time field and scalar parameters, consistent with the data and the model assumptions. The results presented here are based on 2000 such draws, after first running the algorithm for several thousand iterations to allow the chain to 'burn in' [Gelman et al., 2003].

Details of the data and results can be found in the figure captions. Figure 1 describes the data set; Figure 2 shows estimates of the field for six years; Figure 3 shows estimates of time series at two locations; Figure 4 shows the time series of the spatial average; Figure 5 shows the prior and posterior distributions for the scalar parameters; Figure 6 shows selected time series of innovations and observational errors; Figure 7 shows the correlation between the proxy time series and the posterior estimates of temperature time series; Tables 1 and 2 show the proportion of ensemble members for which particular years of the spatial mean time series are most extreme, for the 1400-1999 interval; Figure 8 shows the proportion of ensemble members for which particular years in the 1400-1999 interval are most extreme, as a function of location.

References

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.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman&Hall/CRC, Boca Raton, 2 edition, 2003.

M.P. Tingley and P. Huybers. A Bayesian Algorithm for Reconstructing Spatially Arrayed Part 1: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 2009a.

M.P. Tingley and P. Huybers. A Bayesian Algorithm for Reconstructing Spatially Arrayed Part 2: Behavior and comparison with RegEM. *Journal of Climate*, 2009b.

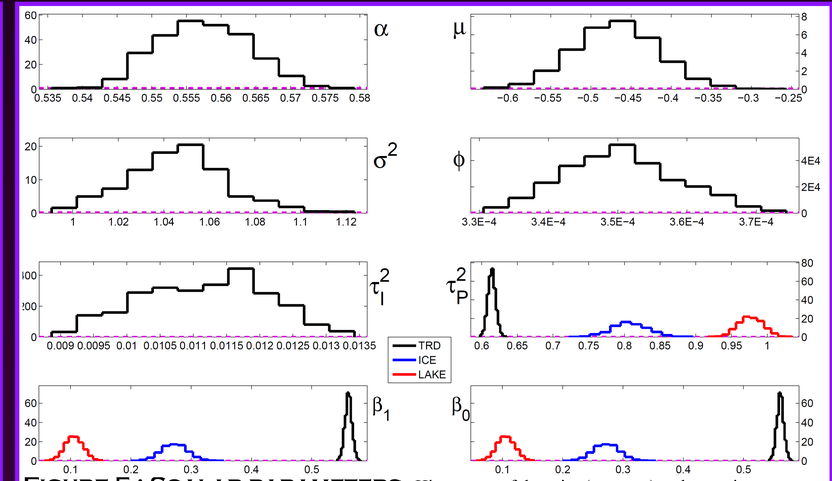

Figure 5 : Scalar parameters. Histograms of the prior (magenta) and posterior distributions of the scalar parameters. Note that the priors are all flat relative to the posteriors. τ^2_P is smallest and β_1 largest for the TRD, indicating that this proxy has the strongest relationship with the underlying field. The different values inferred for τ^2_P and the β's for the different proxy types indicate the need to model them separately.

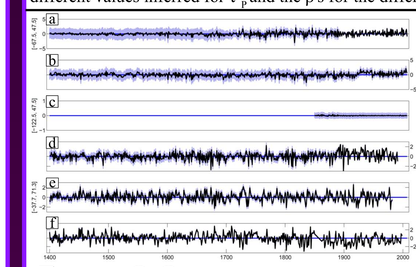

Figure 6: Observational errors and innovations. (a, b) Medians and 90% CIs for the innovation (ϵ_t) time series located in (New Brunswick, Greenland). **(c,d,e,f)** Medians and 90% CIs for representative (instrumental, TRD, ICE, LAKE) observational error time series.

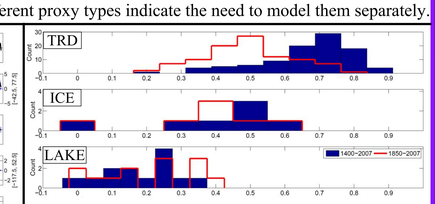

Figure 7 : Correlations. Histograms of the correlation between each proxy time series and the estimated temperature time series at the corresponding location, over the length of the reconstruction and confined to the instrumental interval. Note that the TRD correlations are highest, indicating that these proxy observations are weighted more heavily than the ICE or LAKE observations. TRD correlations are lower over the 1850-2007 interval when the available instrumental observations dominate the temperature estimates.

Table 1: Warm years and decades, 1400-1999.

Region	1995	1998	1990-1999
Land (45N-85N)	0.977	0.000	1.000
NA/GL (45N-85N)	0.000	0.873	0.927
Eurasia (45N-85N)	0.996	0.000	1.000

The proportion of ensemble members for which 1995 and 1998 are the warmest years, and the 1990s the warmest decade, in the 1400-1999 interval. We consider three spatial averages, each confined to the 45°N-85°N latitude band: All land, North America/Greenland, and Eurasia.

Region	1453	1601	1642	1643	1817
Land (45N-85N)	0.048	**0.701**	0.034	**0.152**	0.004
NA/GL (45N-85N)	0.000	**0.367**	0.000	**0.251**	*0.087*
Eurasia (45N-85N)	**0.464**	**0.203**	**0.264**	0.009	0.000

Table 2 : Cold years, 1400-1999. As in Table 1, but for cold years, many of which are associated with major volcanic eruptions (see Figure 4).

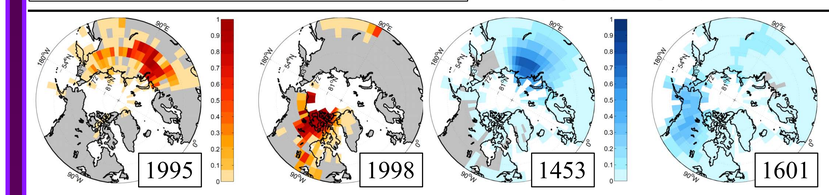

Figure 8: Extreme years, 1400-1999, as a function of space. The proportion of ensemble members for which selected years are the warmest (1995, 1998) or coldest (1453, 1601) in the 1400-1999 interval, as a function of location. The gray shading indicates that at those locations, the selected year is not the most extreme in the 1400-1999 interval for any ensemble member.