Discussion: "The Value of Multi-proxy Reconstruction of Past Climate" by Li, Nychka, and Ammann (LNA)

Noel Cressie (ncressie@stat.osu.edu)

Program in Spatial Statistics and Environmental Statistics

(http://www.stat.osu.edu/∼sses)

The Ohio State University

Joint Research with Martin P. Tingley (SAMSI/NCAR)

Reconstruction of Past Climate

Climate: Temperature, precipitation, air pressure, etc.

Past: Before instrumentation +150 years of instrumentation

Proxy Data: Tree rings, pollen, boreholes, ice cores, etc.

Models: Used to solve a nonlinear, ill-posed inverse problem. Infer climate from proxies that "record" the past environment and from recent instrumental data $\mathbf{T}_2$

In Li, Nychka, and Ammann (**LNA**),

- Climate $\equiv$ Temperature
- Past $\equiv$ 1,150 YBP (Years Before Present)
- Data $\equiv$ Pseudo-proxies (simulated)
- Model $\equiv$ BHM; inverse problem solved using Bayes Theorem
- $\mathbf{T} = \mathbf{T}_0$ (no instrumentation error)

Written discussion of LNA can be found in **Cressie and Tingley (2010)**

Pseudo-proxy (Simulated) Data

- Driven by an atmosphere-ocean general circulation model (**GCM**)
- The true Northern Hemisphere (**NH**) average temperature is given by the GCM: $\mathbf{T}_0 \equiv (\mathbf{T}_{1,0}, \mathbf{T}_{2,0})$
  - $\mathbf{T}_{1,0}$: prior to instrumentation
  - $\mathbf{T}_{2,0}$: temperature **data $\mathbf{T}_2$ available (150 YBP)**
- Pseudo-proxies are functions of $\mathbf{T}_0$; choice of functions should match how proxies depend physically or biologically on temperature
- Think of **pseudo-proxies** as data from **"lab animals,"** and **proxies** are data from **"patients"**
- Science: **Experiments** are conducted on lab animals to infer proper treatment for patients

Experimental Design

Three tenets of **good experimental design** (Fisher, 1935)

- Blocking
- Randomization
- Replication

LNA's Experimental Design

- **Blocking**:

  In LNA's experiment, there are many blocks, representing well thought out factor combinations. But all **blocks are of size one** because there is only **one treatment: Posterior Analysis**. There should be a "status quo" treatment, such as the non-HM analysis called RegEM (Schneider, 2001) that the paleoclimate community use.

  LNA's experiment compares different factor combinations, but does not demonstrate to the paleoclimate community that going to the trouble of building a HM and doing a posterior analysis is worth it. (We believe it is!)

LNA's Experimental Design, ctd.

- **Randomization**:

  Classically, this is used for assignment of treatments to experimental units within a block. Since all blocks in LNA's experiment are of size one, randomization is irrelevant here. (Even if there were another treatment, in a simulation experiment it is possible to apply different treatments to identical experimental units.)

LNA's Experimental Design, ctd.

- **Replication**:

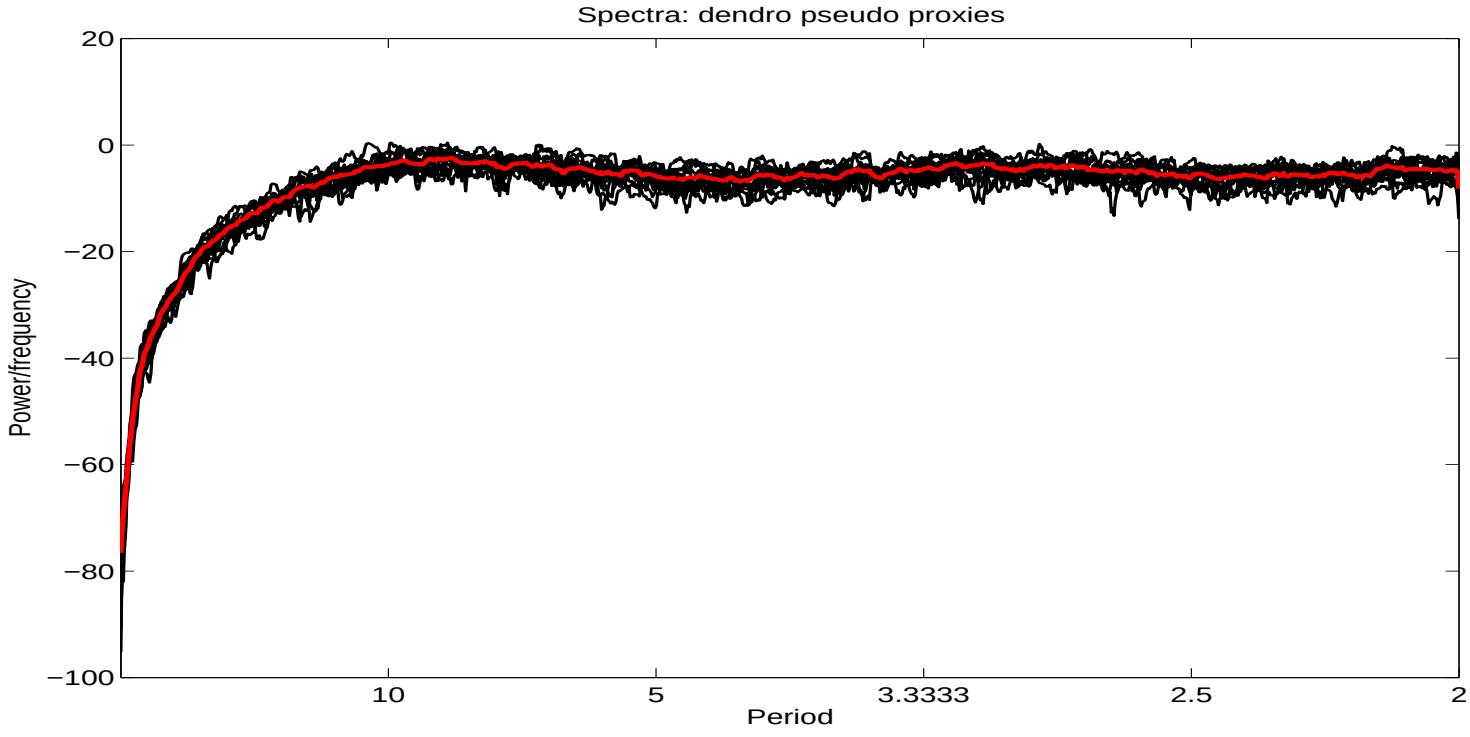
  LNA's experiment has a **sample of size one**, since only **one GCM** is used. That is, there is just one lab animal! It is true there is only one patient (earth's past climate), but that is unknown. So, it would make sense to **replicate the experiment** by choosing at least two other GCMs, giving three replicates (say)

Pseudo-proxy Data: The Lab Animal

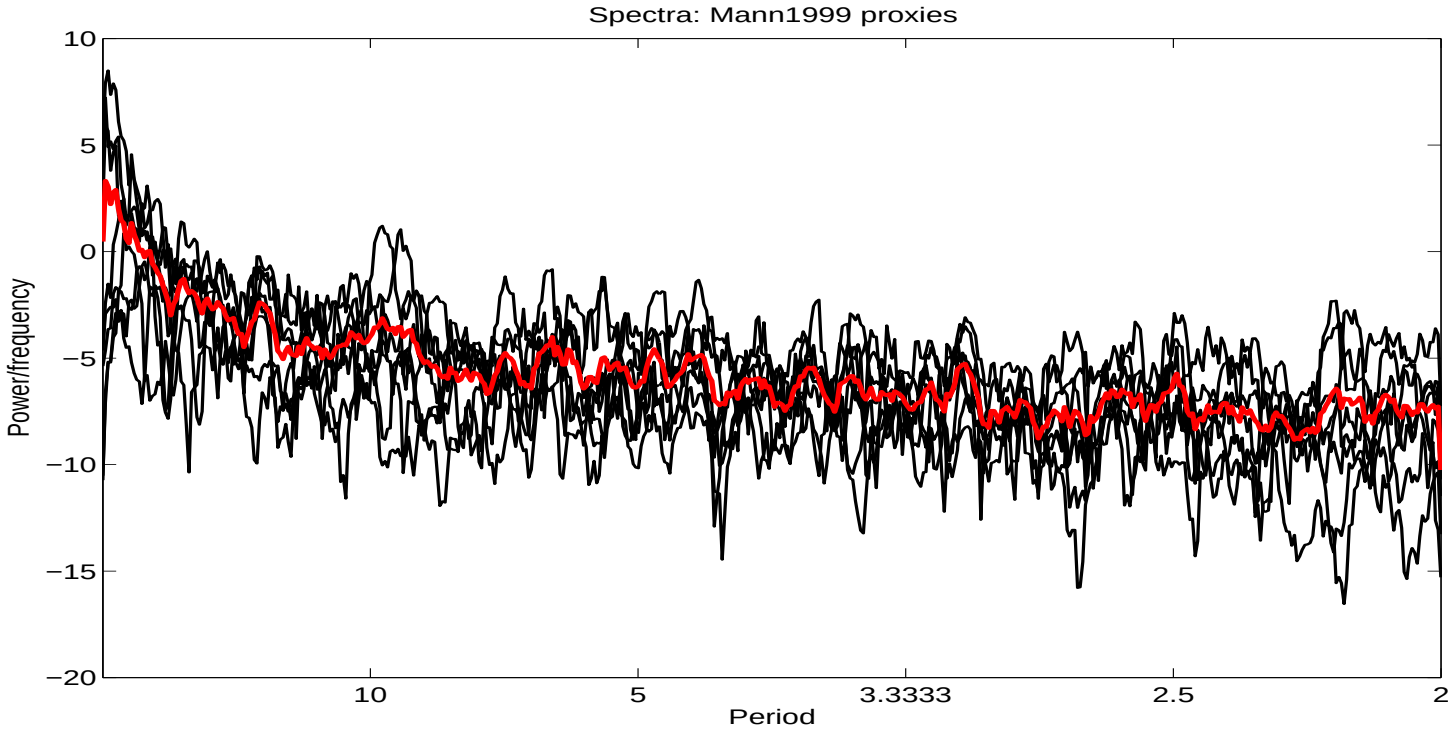
The goal is to make the lab animal like the patient

- **Tree-ring data ($\mathbf{D}$)**: Such data record **high-frequency climatic changes**. LNA mimic this by removing an 11-year running mean from GCM output at a number of locations
  - A spectral analysis of actual tree ring data and the simulated data show the actual data have a longer-range (in time) dependence

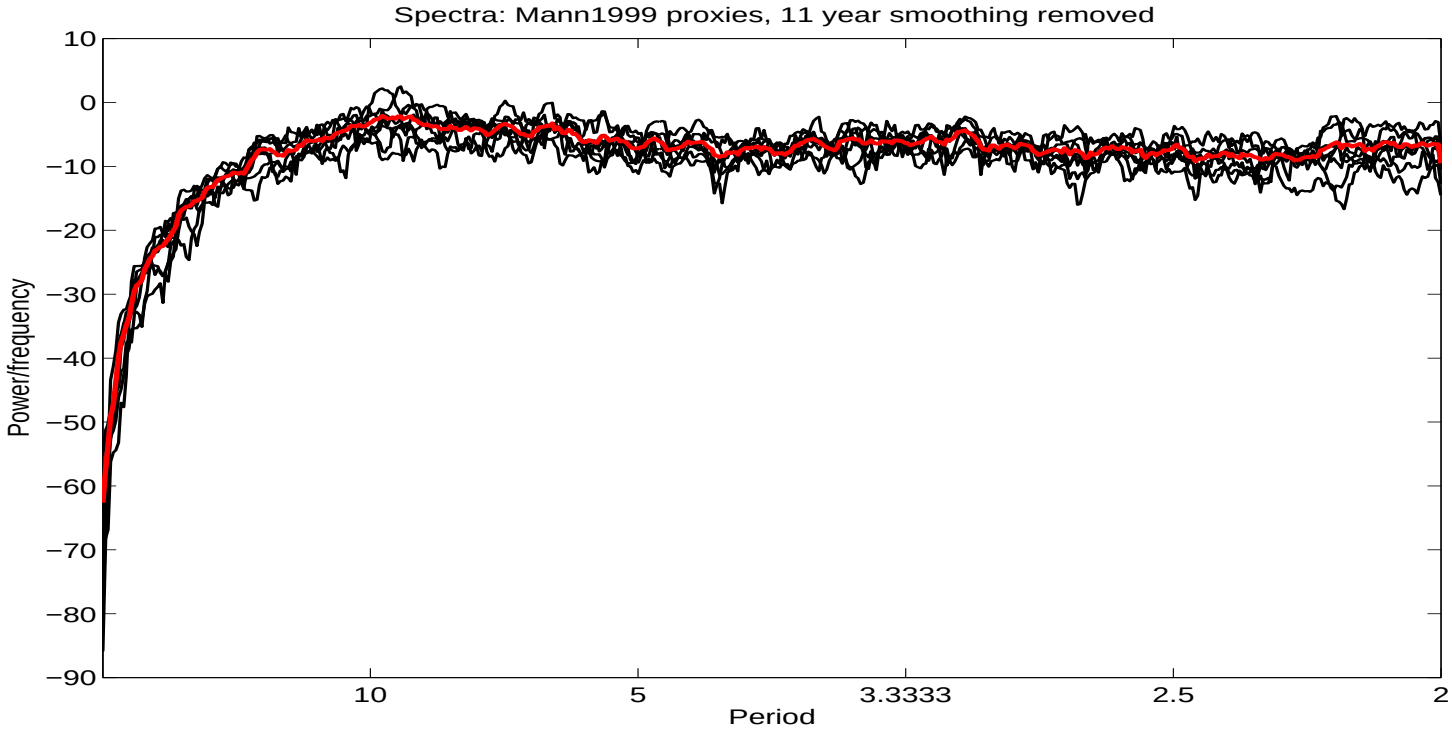
Spectra: Dendro (Tree-ring) Pseudo-Proxies



Spectra: Dendro (Tree-ring) - Mann et al. (1999)



Spectra: Dendro Proxies - 11 Year MA Removed



Pseudo-proxy Data: The Lab Animal

The goal is to make the lab animal like the patient

- **Tree-ring data ($\mathbf{D}$)**: Such data record **high-frequency climatic changes**. LNA mimic this by removing an 11-year running mean from GCM output at a number of locations
  - A spectral analysis of actual tree ring data and the simulated data show the actual data have a longer-range (in time) dependence
  - Use biological models instead of the statistical MA model
  - Choose more appropriate tree locations

Pseudo-proxy Data: The Lab Animal, ctd.

- **Pollen data ($\mathbf{P}$)**: Such data record information about the climate over **large spatial and long temporal scales**. LNA mimic this by averaging GCM output over $7.5° \times 7.5°$ regions and calculating an 11-year running mean
- **Borehole data ($\mathbf{B}$)**: Surface temperature propogates through rock; measurements from a depth profile are the result of diffusion according to the heat equation with surface temperature as the boundary condition. **Data are in temperature units**. LNA mimic this by averaging GCM output over $20° \times 20°$ spatial regions and using the POM-SAT forward model to obtain temperature profiles down to 500m.
  - $20° \times 20°$ regions too large – generous for inferring Northern Hemisphere average temperature
  - These data really require a spatial model for how temperature and proxies are related

The Treatment: Posterior Analysis

Posterior Analysis is based here on a **hierarchical statistical model (HM)**:

- Data model
- Process model
- Parameter model

Generally, “the Science” is in the Process model

Process Model

LNA relate the **(NH average) temperature** time series to **forcings**:

- Solar irradiance ($\mathbf{S}$)
- Volcanism ($\mathbf{V}_0$)
- Greenhouse gases represented by the concentration of CO_2 ($\mathbf{C}$)
- Errors that are not necessarily iid

One of LNA's goals is to **estimate (predict) pre-instrumentation temperature** $\mathbf{T}_{1,0}$ **given** the temperature data $\mathbf{T}_2$, the **(pseudo-)proxies**, and the **forcings**

Process Model, ctd.

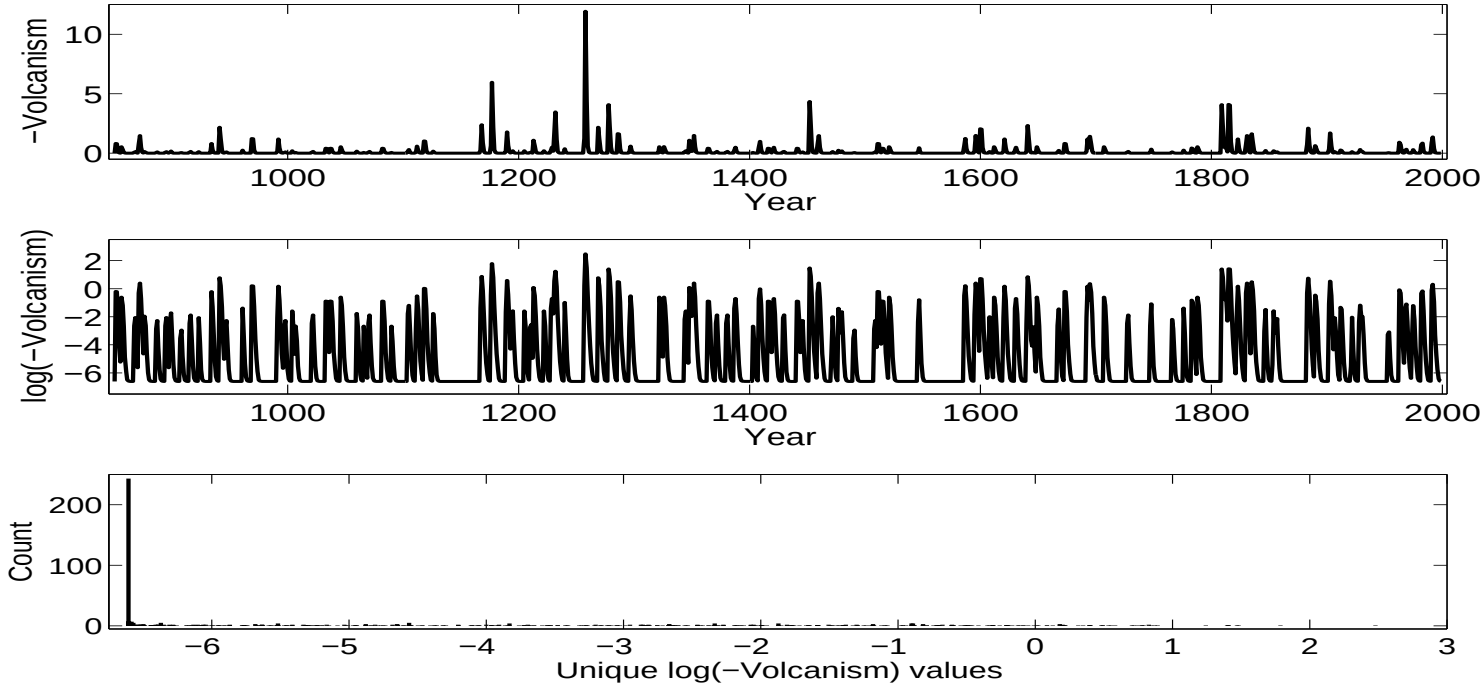
- The basic process model for the 1,150 YBPs is:

$$\mathbf{T}_0 = \beta_0 \mathbf{1} + \beta_1 \mathbf{S} + \beta_2 \mathbf{V}_0 + \beta_3 \mathbf{C} + \boldsymbol{\varepsilon}\,,$$

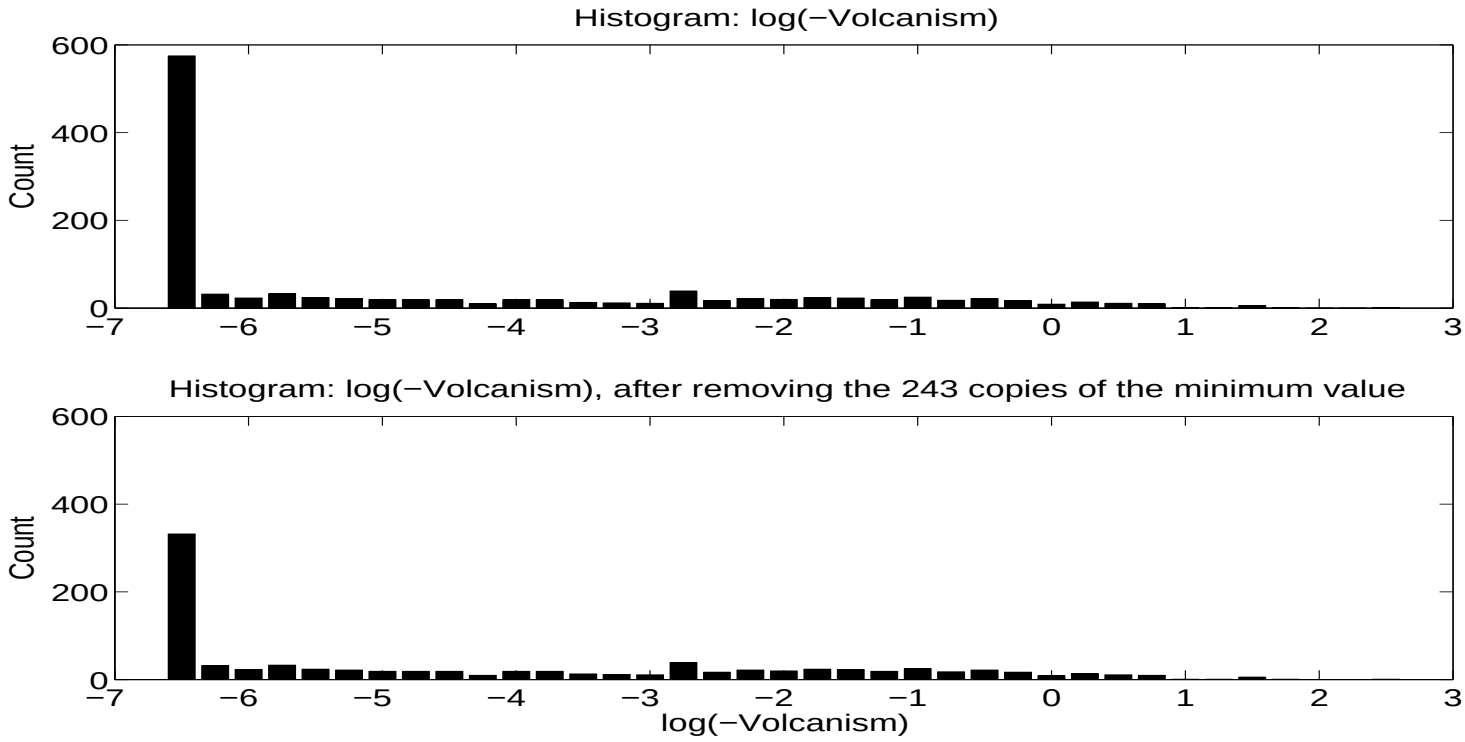
  where $\boldsymbol{\varepsilon} \sim AR(2)$ and all vectors are 1,150-dimensional

  - Data $\mathbf{V}$ are available (modeled conditional on $\mathbf{V}_0$)

Volcanism


−Volcanism
10
5
0
1000
1200
1400
1600
1800
2000
Year
log(−Volcanism)
2
0
−2
−4
−6
1000
1200
1400
1600
1800
2000
Year
Count
200
100
0
−6
−5
−4
−3
−2
−1
0
1
2
3
Unique log(−Volcanism) values

Histograms


Histogram: log(−Volcanism)
Count
600
400
200
0
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
Histogram: log(−Volcanism), after removing the 243 copies of the minimum value
Count
600
400
200
0
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
log(−Volcanism)

Process Model, ctd.

- The basic process model for the 1,150 YBPs is:

$$\mathbf{T}_0 = \beta_0 \mathbf{1} + \beta_1 \mathbf{S} + \beta_2 \mathbf{V}_0 + \beta_3 \mathbf{C} + \boldsymbol{\varepsilon}\,,$$

  where $\boldsymbol{\varepsilon} \sim AR(2)$ and all vectors are 1,150-dimensional

  - Data $\mathbf{V}$ are available (modeled conditional on $\mathbf{V}_0$)
  - $\mathbf{S} = \mathbf{S}_0$, $\mathbf{C} = \mathbf{C}_0$ (no measurement error assumed for solar irradiance and CO_2 concentration)
  - Additive forcings (but perhaps put $\mathbf{V}_0$ and $\mathbf{C}$ on the log scale?)
  - $AR(2)$ errors $\boldsymbol{\varepsilon}$ (but lag 2 seems too small; note that one can invert covariance matrices from an $AR(2)$)

Data Model

Recall proxy data: tree-ring data ($\mathbf{D}$), pollen data ($\mathbf{P}$), and borehole data ($\mathbf{B}$).

- Conditional on $\mathbf{T}_0$,

$$\begin{aligned} \mathbf{D} &= f_D(\mathbf{T}_0; \boldsymbol{\varepsilon}_D) \\ \mathbf{P} &= f_P(\mathbf{T}_0; \boldsymbol{\varepsilon}_P) \\ \mathbf{B} &= f_B(\mathbf{T}_0; \boldsymbol{\varepsilon}_B) \end{aligned}$$

Conditional on $\mathbf{V}_0$,

$$\mathbf{V} = f_V(\mathbf{V}_0; \boldsymbol{\varepsilon}_V)$$

  - The functions f_D, f_P, f_B are **assumed known** up to a simple linear regression
  - f_V does not account for thresholding in $\mathbf{V}$
  - There is no Data model for $\mathbf{S}$ and $\mathbf{C}$ (i.e., assume $\mathbf{S} = \mathbf{S}_0$, $\mathbf{C} = \mathbf{C}_0$)

Spatial Variability

The proxy data have **spatial information** that is averaged out in the Posterior Analysis.

- **Spatio-temporal modeling** is suggested in our written discussion (Cressie, Shi, and Kang, 2010; Tingley and Huybers, 2010a, 2010b)
- Incorporating spatial variability allows **spatial sampling design** of proxy data to be addressed
- **Regional temperatures** can be predicted, not just a NH average

Responses

LNA use **bias and mean squared prediction error (MSPE)** as the basic **responses**. Since they conduct a simulation experiment, they know the true pre-instrumentation temperatures $\mathbf{T}_{1,0}$.

- There is an opportunity in this Bayesian analysis to look at other responses, such as **periods of extreme temperatures**
- There is an opportunity to see if the model-based MSPEs match the empirical MSPEs
- Forcings $\mathbf{S}$, or $\mathbf{C}$, or both are important? **Climate skeptics say $\mathbf{S}$, not $\mathbf{C}$!** This could have been addressed as part of the experiment

Conclusions

- The HM allows transparency of the physical and statistical modeling assumptions. Fitting an HM to perform a Posterior Analysis is a big investment that requires a partnership between paleoclimate and statistical scientists
- Posterior Analysis (the “treatment”) looks promising, but does it perform better than RegEM? How does it perform for other GCMs?
- Using **multi**-proxies is very important. They capture **different scales of temporal variability**, all of which are important for accurate and precise paleoclimate reconstruction

References

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

Cressie, N. and Tingley, M.P. (2010). Comment: Hierarchical statistical modeling for paleoclimate reconstruction. *Journal of the American Statistical Association*, forthcoming.

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

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.

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.

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.

Tingley, M.P. and Huybers, P. (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.