Piecing together the past:

Statistical insights into paleoclimatic reconstructions

Peter F. Craigmile

Department of Statistics, The Ohio State University

http://www.stat.osu.edu/~pfc/

This talk introduces the paper of the same name, by

Martin Tingley, Peter Craigmile, Murali Haran, Bo Li,

Elizabeth Mannshardt, and Bala Rajaratnam

To appear in *Quaternary Science Reviews*.

Acknowledgments

- Supported by the Statistical and Applied Mathematical Sciences Institute (SAMSI) and the National Science Foundation.

- Thanks to SAMSI; conversations with:

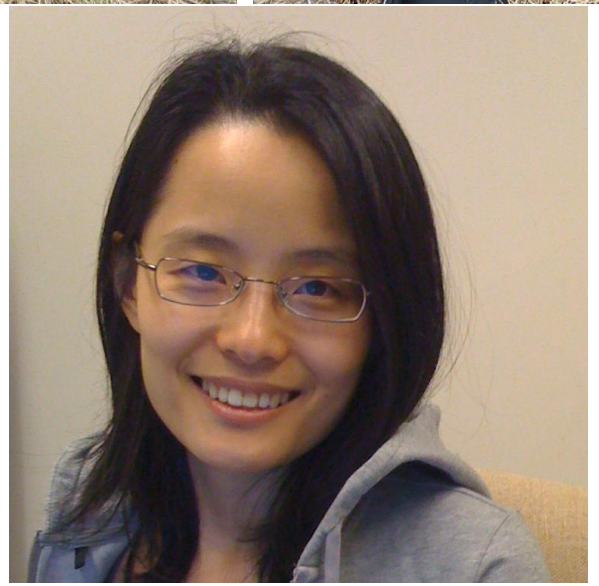
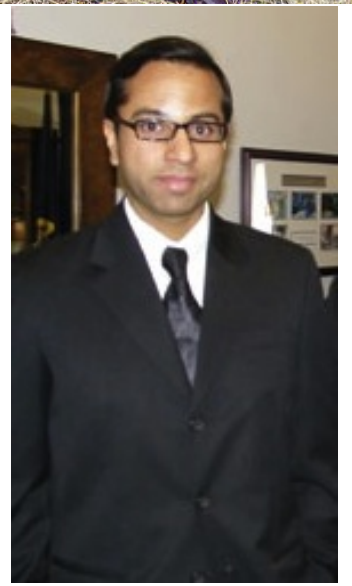
  Bo Christiansen, John Haslett, Cindy Greenwood, Michael Evans, Matthew Schofield,

  and comments from:

  Noel Cressie, Julien Emile-Geay, Michael Mann, Douglas Nychka, Tapio Schneider, Eugene Wahl, and five referees.

The cast



And the cruise director

The SAMSI Paleoclimate working group

- Part of the 2009-2010 SAMSI Program on Space-time Analysis for Environmental Mapping, Epidemiology and Climate Change.
- Our question:

  "What are the **statistical challenges** surrounding the **reconstruction** of **past climate** from incomplete **instrumental** and **proxy** data sets"?
- We believe this is an area where Statistical Scientists can and should collaborate with Paleoclimatologists.

  (Collaborate means "both ways").

The United Kingdom parliamentary report on CRU

- Analyzing climate data has it controversies.

"We cannot help remarking that it is very surprising that research in an area that depends so heavily on statistical methods has not been carried out in close collaboration with professional statisticians. Indeed there would be mutual benefit if there were closer collaboration and interaction between CRU and a much wider scientific group outside the relatively small international circle of temperature specialists."

www.uea.ac.uk/mac/comm/media/press/CRUstatements/SAP

What is our paper not about?

- There is no reconstruction.
- There is no specific statistical analyses of a statistical model.
- We do not test methods of analysis.
- It is not a general review of the climate reconstructions [see, e.g. NRC, 2006, Jones et al., 2009].
- We do not develop statistical models for time-uncertain proxy series [see, e.g. Haslett et al., 2006a, Auestad et al., 2008, Haam and Huybers, 2010].
- We do not discuss paleo data preprocessing [see, e.g., Briffa et al., 1992, Schofield, in prep., Haslett et al., 2006b].

So what do we do?

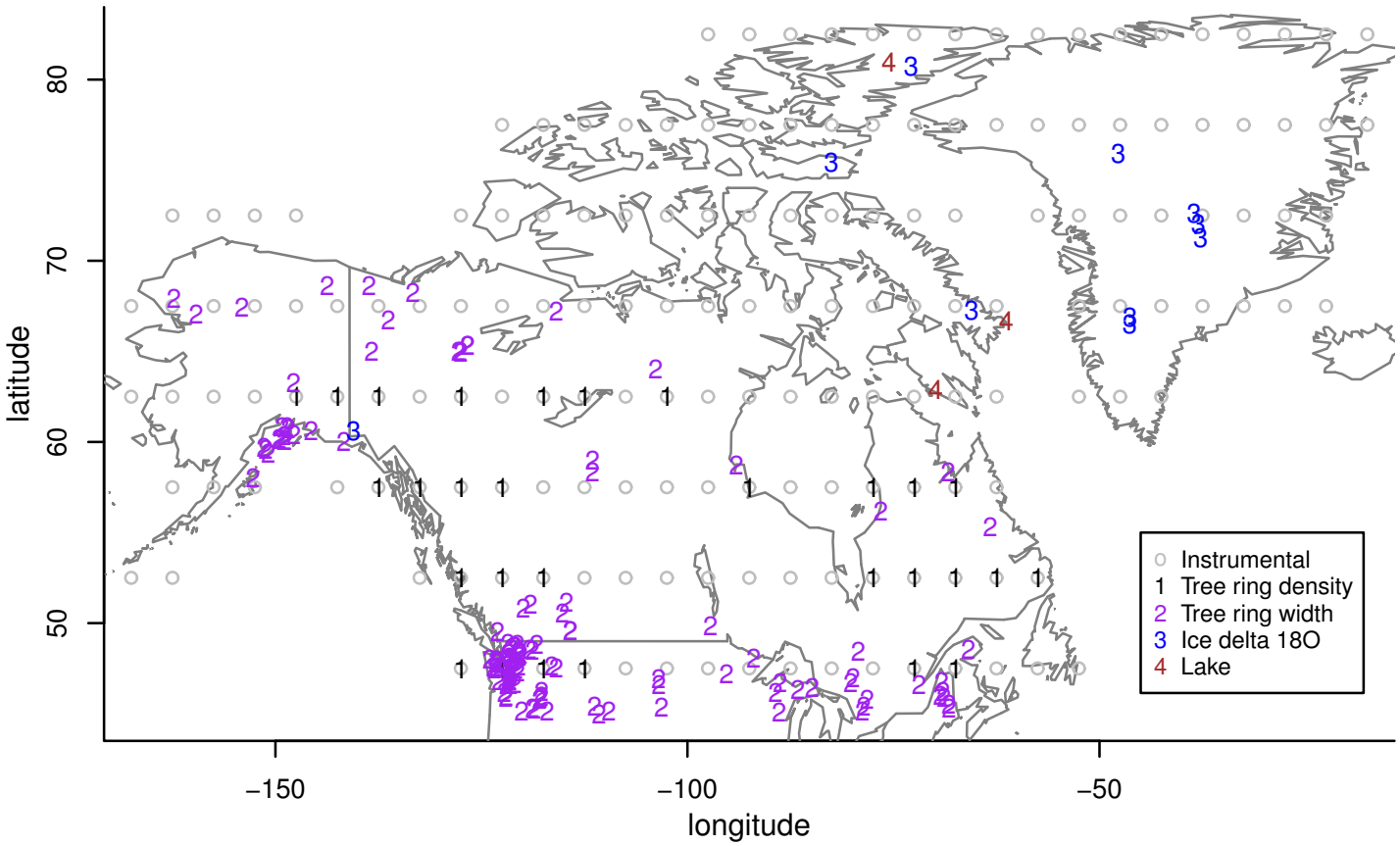
- To demonstrate the role of the **hierarchical statistical models** in (paleo)climate reconstruction problems.
  - To highlight the data-specific, scientific, and statistical modeling challenges in this hierarchical context.
  - To show how existing methods can be interpreted (as close as possible) in this hierarchical context.
- Also see Hughes and Ammann [2009] for an overview of the state of paleoclimate reconstruction methods, and suggestions on how to move forward.

We are not the only statisticians working in the area

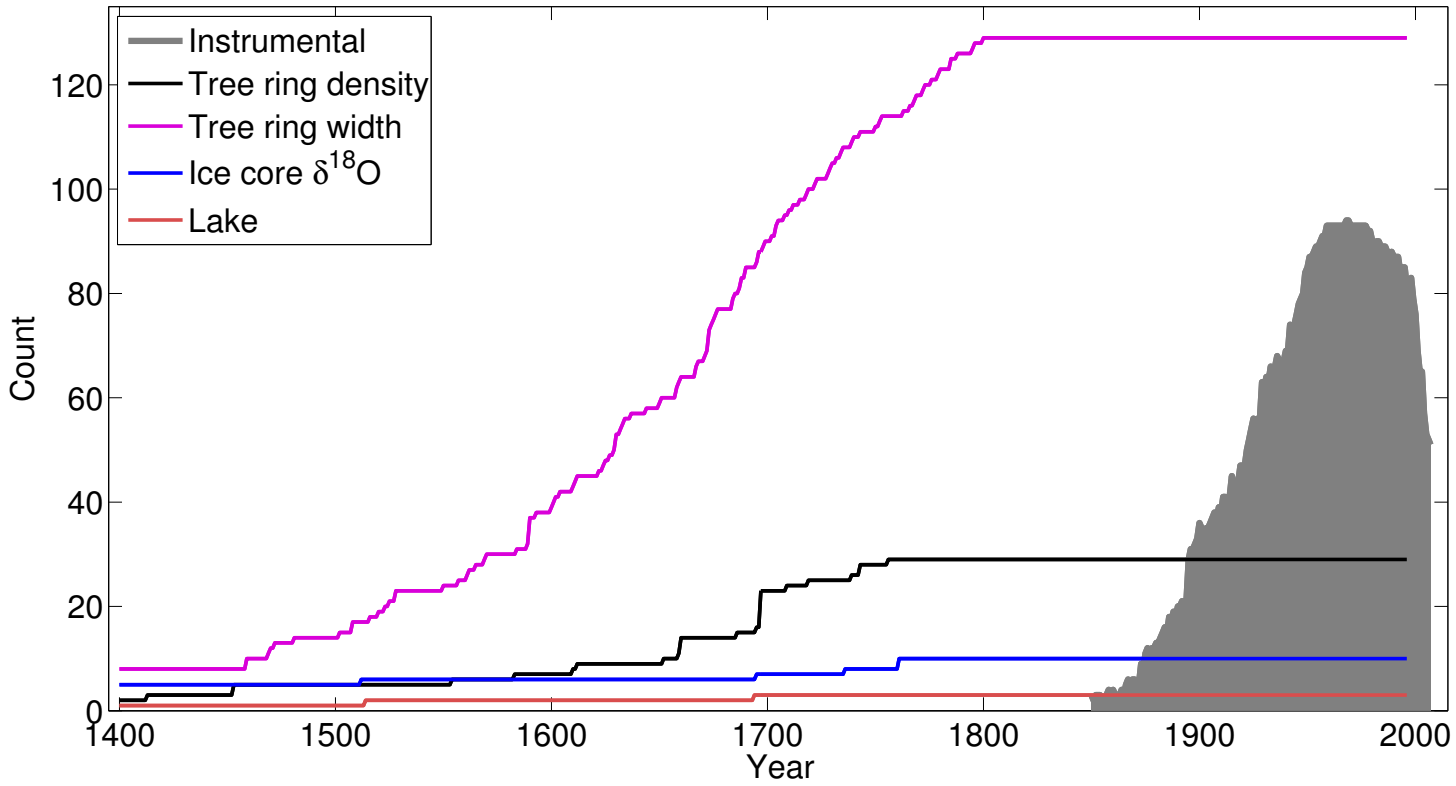
- There are many time series analyses of paleorecords [e.g., West, 1997, Visser and Molenaar, 1988, Harvill and Ray, 2006, Haslett et al., 2006b].
- Li et al. [2010] presents a hierarchical model and applies it to pseudo-proxies derived from climate models.
- Brynjarsdóttir and Berliner [2011] reconstruct surface temperatures using borehole temperature profiles.
- Lee et al. [2008] proposes a state-space or Kalman filter model for inferring large-scale spatial average temperatures.
- Tingley and Huybers [2010a,b] propose a simple hierarchical statistical model without forcings to infer a climate field in both space and time.

Motivation: the Mann et al. [2008] dataset

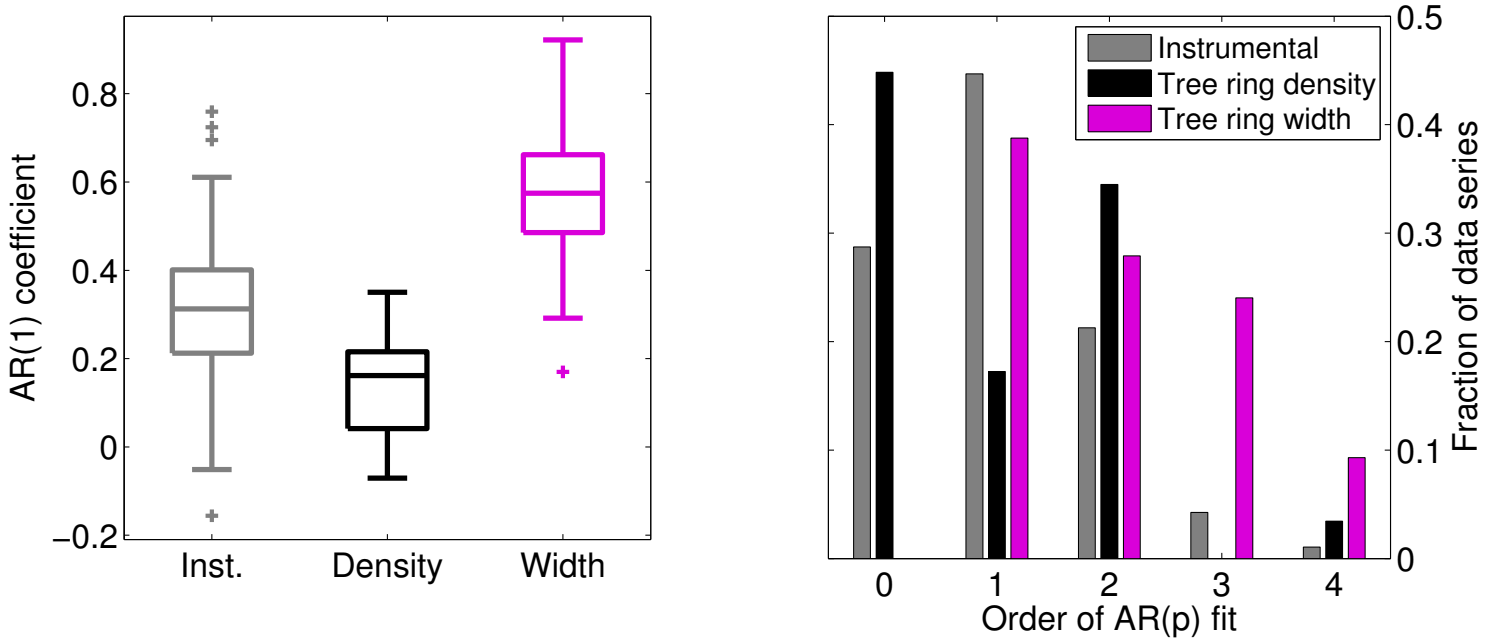
- They reconstruct hemispheric and global surface temperatures over the last two millenia using 1,209 proxy time series and a $5^\circ \times 5^\circ$ gridded surface-temperature data product from CRU.
- Let us focus on Northern North America and Greenland.



A subset of the data used in Mann et al. [2008]. Circles indicate the centroids of the gridded surface-temperature product. Numbers indicate the locations of, or the centroids of the regions represented by, the various proxy time series.



The number of each type of observation as a function of time.



Left panel: Box plots of AR(1) coefficients inferred from the instrumental, tree ring density, and tree ring width time series.

Right panel: optimal order of AR(p) fit, according to the Bayesian Information Criterion [analysis performed with the ARFit Matlab package of Neumaier and Schneider, 2001, Schneider and Neumaier, 2001].

What are we reconstructing?

- A **time series** of large scale spatial averages of the climate field

  [e.g., Moberg et al., 2005, Lee et al., 2008, Mann et al., 2008, Kaufman et al., 2009].

  Composite plus scale (CPS)

- The **spatial pattern** of a climate variable as a **function of time**

  [e.g., Mann et al., 1998, Cook et al., 1999, Luterbacher et al., 2004].

  Climate field reconstruction (CFR)

- A **climate index**, such as El Niño, that reflects broad aspects of climate

  [e.g., Emile-Geay et al., 2012a,b].

A statistical space-time reconstruction

- We focus on the reconstruction of a "target" **latent** climate processes that can be modeled as **continuous in space** and **discrete in time**; for example, annual mean surface temperatures.

  (We'll discuss other domains)

- Different **data sources** may have different **uncertainties** and different **relationships** with the target climate process.

- Each data source, as well as the target process, typically displays **spatial and temporal dependencies**.

The paper layout

(The paper is very long, so read it in **bite-size pieces**!)

3. **Hierarchical statistical models** **(Today's topic)**

4. Modeling the latent space-time climate process

5. Forward models for climate proxies

6. Modeling the observations and other data-level issues

7. Inference and computation

8. Special cases from the literature

9. Discussion

Statistical modeling versus statistical analysis

-

Hierarchical statistical models

- We use **Bayesian inference** as the analysis choice.
- Given specifications of the **prior** distribution, $\pi(\text{parameters})$, for all unknown parameters and the **likelihood**, $f(\text{data}|\text{parameters})$, the **posterior** distribution of the unknown parameters given the data is,

$$\pi(\text{parameters}|\text{data}) \propto f(\text{data}|\text{parameters})\, \pi(\text{parameters}).$$

- cf Noel's talk from last week, we can **decompose** the set of parameters further into **processes** and **parameters**.

A general framework for paleo reconstruction

- We wish to infer upon the **target latent** space-time climate process:

$$\mathbf{Y} = \{Y(\mathbf{s}, t) : \mathbf{s} \in \mathcal{D},\ t \in \mathcal{T}\}$$

- Here $\mathcal{D}$ designates the spatial and $\mathcal{T}$ the temporal domains of interest.
- The form of the domains depend on the spatial and temporal coverages we are interested in.

The data sources: Instrumental observations

- Let

$$\boldsymbol{Z}_{I,j} = \{Z_{I,j}(\mathbf{s},t) \,:\, \mathbf{s} \in \mathcal{D}_{I,j},\ t \in \mathcal{T}_{I,j}\} \qquad j = 1,\ldots,N_I,$$

  denote the N_I different types of instrumental observation, where $\mathcal{D}_{I,j}$ and $\mathcal{T}_{I,j}$ denote the spatial and temporal domains, respectively, for the jth instrumental data type.

- Examples: ground-based thermometers, satellite observations.

- The domains can differ over different data types and from that of $\boldsymbol{Y}$.

The data sources: Proxy records

- Let

$$\boldsymbol{Z}_{P,k} \;=\; \{Z_{P,k}(\mathbf{s},t) \;:\; \mathbf{s} \in \mathcal{D}_{P,k},\; t \in \mathcal{T}_{P,k}\} \qquad k = 1, \ldots, N_P,$$

  denote the N_P different types of proxy records.

- $\mathcal{D}_{P,k}$ and $\mathcal{T}_{P,k}$ are the spatial and temporal domains.

- Examples: the spatially located tree ring density series, tree ring width series, and ice core series in the earlier figure.

Should we just link the data to the target climate field?

- A "cookbook" solution:

  1. Given some parameters, write down the distribution of the data sources given the latent climate process.
  2. Introduce (prior) distributions for the latent climate process and the parameters.
  3. Use Bayes theorem to infer upon the latent climate process and the parameters, conditional on the data.

- Problem with this approach:

  1. Hard to incorporate the science correctly.
  2. Not easy to account for all the different sources of uncertainty.

An example: pollen (p.12)

[See, e.g., Ohlwein and Wahl, 2012]

- Consider a spectrum of **pollen counts** extracted from a sample of a lake sediment core.
- A researcher often extracts a fixed number of grains, sorted by taxa.
- Conditional on the overall count and the probability of a given grain belonging to a taxon, the observed count of the taxon follows a binomial distribution.
- The observed counts are thus used to estimate the parameter of a binomial distribution, and uncertainty is introduced by the limited sample size and effects such as the preferential degradation of certain pollen species.

Pollen continued: linking to climate

- A simple model for the pollen–climate relationship may state that larger proportions of pollen from a particular, **indicator** taxon correspond to warmer temperatures.
- In addition, the **model** relating the pollen counts to the climate is likely an **imperfect representation** of the factors that affect the pollen spectra, in the sense that, given the actual (as opposed to estimated) parameters of the model, there remains uncertainty about the state of the climate system.

Two sources of uncertainty

1. The limitations of the model relating the proxy or instrument to the climate.
2. The limitations of the observations, including measurement errors and finite sample size.

- Motivates a two-stage modeling approach.
  1. We model the distribution of the error-free proxy or instrument process given climate ("the science" or "the forward model").
  2. We model the distribution of the proxy or instrumental data conditional on the error-free proxy or instrument process.

(There is maybe an argument that we should break it down further.)

The error-free instrumental processes

- The N_I latent, error-free, instrumental processes, associated with $\boldsymbol{Z}_{I,j}$:

$$\boldsymbol{W}_{I,j} = \{W_{I,j}(\mathbf{s},t) : \mathbf{s} \in \mathcal{D}_{I,j},\ t \in \mathcal{T}_{I,j}\}, \qquad j = 1,\ldots,N_I$$

- Example: CRU gridded temperature anomaly product.

  "The two-stage model provides flexibility in modeling the key features of the data, including the spatial averaging of the underlying temperature field, the spatially and temporally varying availability of station observations within the grid boxes, and uncertainties associated with the raw station data."

The error-free proxy processes

- The N_P latent, error-free, proxy processes, associated with $\boldsymbol{Z}_{P,k}$:

$$\boldsymbol{W}_{P,k} = \{W_{P,k}(\mathbf{s},t) \ : \ \mathbf{s} \in \mathcal{D}_{P,k},\ t \in \mathcal{T}_{P,k}\}, \qquad k = 1, \ldots, N_P$$

- Example: For pollen,

  $W_{P,1}$ is the true proportion of pollen, and

  $Z_{P,1}$ is the observed pollen spectra.

The likelihood is a product of

1. The joint distribution of the latent space-time climate process $\boldsymbol{Y}$;

2. The joint distribution of the error-free instrumental and proxy processes, $\{\boldsymbol{W}_{I,j} : j = 1 \ldots N_I\}$ and $\{\boldsymbol{W}_{P,k} : k = 1 \ldots N_P\}$, conditional on $\boldsymbol{Y}$;

3. The joint distribution of the instrumental and proxy data, $\{\boldsymbol{Z}_{I,j}\}$ and $\{\boldsymbol{Z}_{P,k}\}$, conditional on the error-free processes $\{\boldsymbol{W}_{I,j}\}$ and $\{\boldsymbol{W}_{P,k}\}$ and the climate process $\mathbf{Y}$.

We also introduce priors and covariates

- **Parameters** $\boldsymbol{\theta}$: a number of unknown statistical parameters (such as autoregressive coefficients, spatial ranges, and measurement error variances). We need to specify a prior distribution for $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$.
- **Covariates**: such as latitude, longitude, proximity to a coastline, or spatial maps indicating where trees grow over the globe.

The posterior

- **Assume** that the measurement error mechanisms are conditionally independent across data sources, and do not depend on the climate process $\mathbf{Y}$.
- Then the, posterior distribution is

$$\begin{aligned}
&\pi(\mathbf{Y}, \{\mathbf{W}_{I,j}\}, \{\mathbf{W}_{P,j}\}, \boldsymbol{\theta} \mid \{\mathbf{Z}_{I,j}\}, \{\mathbf{Z}_{P,k}\}) \\
&\quad \propto f(\mathbf{Y}|\boldsymbol{\theta})\, g\left(\{\mathbf{W}_{I,j}\}, \{\mathbf{W}_{P,k}\}|\mathbf{Y}, \boldsymbol{\theta}\right) \\
&\qquad \times \left[\prod_{j=1}^{N_I} h_{I,j}(\mathbf{Z}_{I,j}|\mathbf{W}_{I,j}, \boldsymbol{\theta})\right] \left[\prod_{k=1}^{N_P} h_{P,k}(\mathbf{Z}_{P,k}|\mathbf{W}_{P,k}, \boldsymbol{\theta})\right] \pi(\boldsymbol{\theta}).
\end{aligned}$$

- We can use Markov chain Monte Carlo (MCMC) to draw samples from this posterior.

The devil is in the details: for later weeks

4. Modeling the latent space-time climate process

5. Forward models for climate proxies

6. Modeling the observations and other data-level issues

7. Inference and computation

8. Special cases from the literature

9. Discussion

References

B. H. Auestad, R. H. Shumway, D. Tjøstheim, and K. L. Verosub. Linear and nonlinear alignment of time series with applications to varve. *Environmetrics*, 19:409–427, 2008.

K. Briffa, P. Jones, T. Bartholin, and D. Eckstein. Fennoscandian summers from AD 500: temperature changes on short and long timescales. *Climate Dynamics*, 7(3):111–119, 1992.

J. Brynjarsdóttir and L. M. Berliner. Bayesian hierarchical modeling for paleoclimate reconstruction from geothermal data. *The Annals of Applied Statistics*, 5(2B):1328–1359, 2011.

E. Cook, D. Meko, D. Stahle, and M. Cleaveland. Drought reconstructions for the continental united states. *Journal of Climate*, 12:1145–1162, 1999.

J. Emile-Geay, K. Cobb, M. Mann, S. Rutherford, and A. T. Wittenberg. Estimating tropical pacific SST variability over the past millennium. Part 1: Methodology and validation. *Journal of Climate*, 2012a. Submitted; currently available at: http://college.usc.edu/labs/jeg/publications/.

J. Emile-Geay, K. Cobb, M. Mann, S. Rutherford, and A. T. Wittenberg. Estimating tropical pacific SST variability over the past Millennium. Part 2: Reconstructions and uncertainties. *Journal of Climate*, 2012b. Submitted; currently available at: http://college.usc.edu/labs/jeg/publications/.

E. Haam and P. Huybers. A test for the presence of covariance between time-uncertain series of data with application to the Dongge Cave speleothem and atmospheric radiocarbon records. *Paleoceanography*, 25(2):PA2209, 2010.

J. Harvill and B. Ray. Functional coefficient autoregressive models for vector time series. *Computational Statistics and Data Analysis*, 50(12): 3547–3566, 2006.

J. Haslett, A. Parnell, and M. Salter-Townsend. Modelling temporal uncertainty in palaeoclimate reconstructions. In *Proceedings of the 21st International Workshop on Statistical Modelling*, pages 26–37, 2006a.

J. Haslett, M. Whiley, S. Bhattacharya, M. Salter-Townshend, S. Wilson, J. Allen, B. Huntley, and F. Mitchell. Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):395–438, 2006b.

M. Hughes and C. Ammann. The future of the past – an earth system framework for high resolution paleoclimatology: editorial essay. *Climatic Change*, 94(3):247–259, 2009.

P. Jones, K. Briffa, T. Osborn, J. Lough, T. van Ommen, B. Vinther, J. Luterbacher, E. Wahl, F. Zwiers, M. Mann, et al. High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene*, 19(1):3–49, 2009.

D. Kaufman, D. Schneider, N. McKay, C. Ammann, R. Bradley, K. Briffa, G. Miller, B. Otto-Bliesner, J. Overpeck, B. Vinther, et al. Recent warming reverses long-term arctic cooling. *Science*, 325(5945):1236, 2009.

T. Lee, F. Zwiers, and M. Tsao. Evaluation of proxy-based millennial reconstruction methods. *Climate Dynamics*, 31(2):263–281, 2008.

B. Li, D. Nychka, and C. Ammann. The value of multi-proxy reconstruction of past climate. *Journal of the American Statistical Association*, 105 (491):883–911, 2010.

J. Luterbacher, D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner. European seasonal and annual temperature variability, trends, and extremes since 1500. *Science*, 303(5663):1499–1503, 2004.

M. Mann, R. Bradley, and M. Hughes. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, 392(6678):779–787, 1998.

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):13252–13257, 2008.

A. Moberg, D. Sonechkin, K. Holmgren, N. Datsenko, and W. Karlén. Highly variable northern hemisphere temperatures reconstructed from low-and high-resolution proxy data. *Nature*, 433(7026):613–617, 2005.

A. Neumaier and T. Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):27–57, 2001.

NRC. *Surface Temperature Reconstructions for the Last 2000 Years*. The National Academies Press, Washington, D.C., 2006.

C. Ohlwein and E. R. Wahl. Review of probabilistic pollen-climate transfer methods. *Quaternary Science Reviews*, 2012. in press.

T. Schneider and A. Neumaier. Algorithm 808: Arfit—a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):58–65, 2001.

M. Schofield. Climate reconstruction using tree-ring data. in prep.

M. 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. 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), 2010b.

H. Visser and J. Molenaar. Kalman filter analysis in dendroclimatology. *Biometrics*, 44(4):929–940, 1988.

M. West. Time series decomposition. *Biometrika*, 84(2):489, 1997.