Application of Statistical Approaches in Past Temperature Reconstruction

Bo Li, Douglas Nychka
and
Caspar Ammann

National Center for Atmospheric Research

Supported by the National Science Foundation
PART I

Uncertainties in Past Temperature Reconstruction

- Ensemble reconstruction and Bayesian Reconstruction
Why care about the PAST temperature?

- **Long time series** of climate variables including temperature are required to understand the dynamics of climate change
- **Direct observations** of surface temperature is only available from 1850

**How to get past temperatures?**

Reconstruct the past temperature from indirect observations (**proxies**) such as **Tree Ring**, **Pollen** and **Borehole** and **Radiative Forcings**
A stats perspective

The problem:

Quantify the uncertainty in the temperature estimates from other kinds of observations (i.e proxies).

e.g. What is the uncertainty in the maximum decadal temperature estimates for the last 1000 years?

A statistical solution:

Find the distribution of the temperatures given the other observations.

Represent this distribution by an ensemble of possible reconstructions all statistically valid.
Exploit where there is overlap in two sets of data.
A minimal recipe for generating ensembles:

- A **stationary linear** relationship between the temperature and proxies
- The conditional distribution of temperature given proxies is **normally** distributed.
- Prediction errors are **correlated in time**
- **Adjustment made for overfitting** during calibration period using cross-validation
- Also include the **uncertainty in the parameters**.
Linear prediction of the expected temperature based on proxies.
Linear prediction of the expected temperature based on proxies.

Clearly the errors are correlated.
Linear Model of NH temperature on proxies

The statistical model:

\[ T_t = \mathbf{p}_t' \beta + e_t \]

\( T_t, \mathbf{p}_t \): the temperature and proxies at time \( t \).
\( \beta \): a vector of regression coefficients

The vector of error \( e_t \) follows an AR(2) process:

\[ e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \epsilon_t, \quad \epsilon_t \sim \text{iid Normal}(0, \sigma^2) \]

Model fitting procedure: Generalized Least Squares
Symptom of overfitting: The variance of observed prediction errors is greater than the prediction variability derived from the model.

Solution: 10-fold cross-validation to estimate the inflation adjustment

Uncertainty of model parameter estimates:
Parametric bootstrap to estimate the sampling distribution of the parameter estimates.

An ensemble:

\[ \tilde{T} = P\tilde{\beta} + (e_{1000}, ..., e_{1849})'(e_{1850}, ..., e_{1980})' \]
A draw from the conditional distribution of temperature given the values of the proxies.
Reconstructed temperatures

A second ensemble member

A draw from the conditional distribution of temperature given the values of the proxies.
A draw from the the conditional distribution of temperature given the values of the proxies.
Reconstructed temperatures

Decadal means – three ensembles
Reconstructed temperatures

95 % uncertainty bounds decadal means

Degree C

1000 1150 1300 1450 1600 1750 1900
−0.6 −0.3 0.1 0.4

Degree C

1000 1150 1300 1450 1600 1750 1900
−0.6 −0.3 0.1 0.4
Reconstructed temperatures

Maxima uncertainty
Reconstructed temperatures

Maxima uncertainty using box plots

[Graph showing box plots and time series data with temperature on the y-axis and time on the x-axis.]
Reconstructed temperatures

Maxima uncertainty 95% upper bound
Reconstructed temperatures

The Program:

• Develop a statistical (perhaps complicated) relationship between temperatures and proxies.

• The prediction is the distribution of the temperatures given the observed values of the proxies.

What can not be addressed:

Errors in the proxies and analysis outside the statistical model. (We don’t know what we don’t know.)

e.g. Proxies change in their relationship to temperature over time
Hierarchical Bayesian Model (HBM)

Three hierarchies:

- **Data Stage:** [Proxies|Temperature, Parameters]
  Likelihood of Proxies given temperatures
- **Process Stage:** [Temperature|Parameters]
  Physical model of temperature process
- **Parameter Stage:** [Parameters]
  Specify the prior of parameters
An example of HBM: simulated numerical data
Output from global coupled climate model

pseudo-proxy series sampled from 14 individual grid boxes in the climate model
process model:

**Radiative forcings**: Explosive volcanism, Solar Activity Changes and Anthropogenic forcings
Process Model

\[(\mu_{01}, \mu_{02})^\top: \text{Radiative forcings}\]

Let

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 1 & \mu_{01} \\ 1 & \mu_{02} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}
\]

\[(T_1^\top, T_2^\top)|\mu^\top, \phi_1, \phi_2, \sigma_T^2 \sim MN(\mu^\top, \Sigma_{(\phi_1, \phi_2, \sigma_T^2)})\]

- \(\mu_{01}\): External forces before 1850
- \(\mu_{02}\): External forces 1850-1980

- \(\phi_1, \phi_2, \sigma_P^2\): first and second order time lag coefficient and variance parameter of the AR(2) model
Simulated numerical data

Input: 14 pseudo-proxies and full model temperature

1854-1980
Reconstruct the past temperature (850-1853)
Conclusion from the example

- The posterior mean matches the trend of the numerical data
- The numerical data is within the 95% prediction band
- The posterior mean of parameters are close to those directly estimated from numerical data

HBM works well in reconstructing the past temperature
Revisit the MBH 99 data using HBM

Input: 14 proxies and the instrumental temperature
Reconstruct the past temperature (1000-1849)
Reconstruct real world NH temperature

The period with instrumental records.
Reconstruct real world NH temperature

Maxima uncertainty 95% upper bound
Combining Information from Different Sources

– New Application of Bayesian Hierarchical Models
Data - **Tree Ring, Pollen and Borehole**
Data - Tree Ring, Pollen and Borehole
Energy Balance Model

An object will warm or cool depending on its energy imbalances

Three forcings for global climate system:

1. **Solar radiation**: A positive radiative forcing tends to warm the surface on average, whereas a negative one tends to cool it

2. **Volcanism**: block the solar radiation due to the large amounts of aerosols ejected by volcanic eruption into the atmosphere

3. **Greenhouse gases**: absorb infrared radiation, trap heat within the atmosphere.
Forcings

a: **Volcanism** (contains substantial noise)
b: **solar radiance**
c: **green house gases**
Formulate the problem

Skill of each proxy and forcings

- Tree ring: annual to decadal
- Pollen: bi-decadal to semi-centennial
- Borehole: centennial and onward
- Forcings: external drivers

Goal: Reconstruct the 850-1849 temperature by all proxies, forcings and the 1850-1999 temperature

Bayesian Hierarchical Model (BHM) to thread all proxies, forcings and temperatures
Bayesian Hierarchical Model (BHM)

Bayes’ theorem

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

\( P(A|B) \): the posterior probability because it is derived from or depends upon the specified value of \( B \).

\( P(B) \): the prior or marginal probability of \( B \), and acts as a normalizing constant.

Intuitively, Bayes’ theorem describes the way in which one’s beliefs about observing ’\( A \)’ are updated by having observed ’\( B \)’.
A natural framework: Three hierarchies:

- **Data Stage:** [Proxies | Geophysical Process, Parameters]
  Likelihood of Proxies given underlying process

- **Process Stage:** [Geophysical Process | Parameters]
  Physical model of Geophysical Process

- **Parameter Stage:** [Parameters]
  Specify the prior of parameters
Data Stage: [Data|Geophysical Process, Parameters]

Forward models for tree ring ($R$) and pollen ($P$):

$$R_i|(T_1, T_2) = \beta_{01i} + \beta_{11i}M_1(T_1, T_2) + \epsilon_{1i}, \quad \epsilon_{1i} \sim \text{AR}(2)$$

$$P_i|(T_1, T_2) = \beta_{02i} + \beta_{12i}M_2(T_1, T_2) + \epsilon_{2i}, \quad \epsilon_{2i} \sim \text{AR}(2)$$

$$\epsilon_{1i} \sim \text{AR}(2)(\sigma_1^2, \phi_{11}, \phi_{21}); \quad \epsilon_{2i} \sim \text{AR}(2)(\sigma_2^2, \phi_{12}, \phi_{22})$$

$T_1$: unknown temperature for the past millenium

$T_2$: observed temperature

$M_1$ and $M_2$: transformation matrices for tree ring and pollen, respectively
Tree Ring and Pollen:

Blue: temperature; Black: tree ring;
Red: Pollen, observed at every 30 years
Data Stage: Forward models for Borehole ($B$):

$$B_i|(T_1, T_2) = M_3\{\beta_{03i} + \beta_{13i}(T_1, T_2) + \epsilon_{3i}\}, \quad \epsilon_{3i} \sim \text{iid Normal}$$

$$\epsilon_{3i} \sim \text{iid } N(0, \sigma_{3}^2)$$

$M_3$: transformation matrix for borehole

- Developed from pre-observation mean–surface air temperature (POM–SAT) model based on heating equation
- Respect the smoothness of depth profile
**Data Stage:**

**Volcanism** contain considerable noise among three forcings

\[ V|V' = (1 + \epsilon_4)V', \quad \epsilon_4 \sim \text{iid } N(0, 1/64) \]

*V*: **observed volcanism**

*V’*: **ideal volcanism without noise**

The magnitude of noise in *V* depends on mean *V’*
**Process Stage:** [Geophysical Process|Parameters]

\[(T_1, T_2) | (S, V', C) = \beta_0 + \beta_1 S + \beta_2 V' + \beta_3 C + \epsilon_5, \quad \epsilon_5 \sim \text{AR}(2)\]

\[\epsilon_5 \sim \text{AR}(2)(\sigma^2_5, \phi_{13}, \phi_{23})\]

\(S:\) solar radiance

\(V':\) Volcanism

\(C:\) green house gases, mainly include \(CO_2\), methane

Can be replace by more complicated dynamic model, e.g., General Circulate Model (GCM)
Conjugate priors allow for an explicit full conditional posterior distribution

\[ \beta_{jk_i} \sim N(\tilde{\mu}_{jk_i}, \tilde{\sigma}^2_{jk_i}) \quad j = 0, 1; \quad k = 1, 2, 3, \]
\[ \beta_i \sim N(\tilde{\mu}_i, \tilde{\sigma}^2_i) \quad i = 0, 1, 2, 3, \]
\[ \sigma^2_i \sim IG(\tilde{q}_i, \tilde{r}_i) \quad i = 1, \ldots, 5 \]

Guarantees their corresponding AR process to be stationary and causal (Shumway and Stoffer, 1999, ch. 3)

\[ \phi_{2i} \sim \text{unif}(-1, 1), \quad \phi_{1i}|\phi_{2i} \sim (1 - \phi_{2i}) \times \text{unif}(-1, 1) \quad i = 1, 2, 3, \]
Posterior distribution:

\[
[T_1, \beta_{01i}, \beta_{11i}, \beta_{02i}, \beta_{12i}, \beta_{03i}, \beta_{13i}, \beta_i, \phi_{1i}, \phi_{2i}, \sigma_i^2, V'|R_i, P_i, B_i, S, V, C, T_2]
\]

\[\propto [R_i|(T_1, T_2)|\beta_{01i}, \beta_{11i}, \phi_{11}, \phi_{21}, \sigma_1][P_i|(T_1, T_2)|\beta_{02i}, \beta_{12i}, \phi_{12}, \phi_{22}, \sigma_2][B_i|(T_1, T_2)|\beta_{03i}, \beta_{13i}, \sigma_3][(T_1, T_2)|(S, V', C)][V|V']\]

One example:

\[
[T_1|\bullet] = N(\text{Ab, A})
\]

By completing square:

\[
A = (\sum_{k=1}^{3} \beta_{1ki}^2 M_k^T V_k^{-1} M_k + V_T^{-1})^{-1}
\]

\[
b = \sum_{k=1}^{3} \beta_{1ki} M_k^T V_k^{-1} (\text{Proxy}_k - \beta_{0ki}) + V_T^{-1} (\beta_0 + \beta_1 S + \beta_2 V' + \beta_3 C)
\]
Normal and Inverse Gamma:

**Inverse Gamma:**

\[ f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{1}{x\beta}} \]

\[ E(x) = \frac{1}{(\alpha - 1)\beta} \]

\[ \text{var}(x) = \frac{1}{(\alpha - 1)^2 (\alpha - 2)\beta^2} \]

**Multivariate Normal:**

\[ f(y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)} \]  

(1)
If $\Sigma$ can be written as $\sigma^2 \tilde{\Sigma}$, i.e., a product of constant variance $\sigma^2$ and correlation matrix $\tilde{\Sigma}$, (1) becomes into:

$$f(y; \mu, \sigma^2, \tilde{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \sigma^2 n |\tilde{\Sigma}|}} e^{-\frac{1}{2\sigma^2}(y-\mu)^T \tilde{\Sigma}^{-1}(y-\mu)}.$$  

Suppose $\sigma^2 \sim IG(q, r)$,

$$[y|\mu, \sigma^2, \tilde{\Sigma}][\sigma^2] \propto (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y-\mu)^T \tilde{\Sigma}^{-1}(y-\mu)} (\sigma^2)^{-(q+1)} e^{-\frac{1}{r\sigma^2}}$$

$$= (\sigma^2)^{-(q+n/2+1)} e^{-\frac{1}{\sigma^2}\left\{\frac{1}{r} + \frac{(y-\mu)^T \tilde{\Sigma}^{-1}(y-\mu)}{2}\right\}}$$

So the posterior $[\sigma^2|\cdot]$ is

$$IG(q + n/2, \left\{1/r + 0.5(y - \mu)^T \tilde{\Sigma}^{-1}(y - \mu)\right\}^{-1}$$
MCMC

generate posteriors by alternating Gibbs sampler and Metropolis-Hasting algorithm

1. $T_1 | \cdot$
2. $P|\cdot, B|\cdot, S|\cdot$
3. $(\beta_{01}, \beta_{11})|\cdot, (\beta_{02}, \beta_{12})|\cdot, (\beta_{03}, \beta_{13})|\cdot, (\beta_0, \beta_1, \beta_2, \beta_3)|\cdot$
4. $\sigma^2_1|\cdot, \sigma^2_2|\cdot, \sigma^2_3|\cdot, \sigma^2_4|\cdot, \sigma^2_5|\cdot, \sigma^2_6|\cdot, \sigma^2_t|\cdot$
5. $(\phi_{1r}, \phi_{2r})|\cdot, \phi_p|\cdot, \phi_b|\cdot, (\phi_{1t}, \phi_{2t})|\cdot$ (M-H Algorithm because the conditional distribution has no closed form)

Scrambling the order does NOT affect the result
Implement BHM using Climate from Models

Generate pseudo proxies:

**15 tree rings**: remove **10** year smoothing average

**10 pollen**: **10** year smoothing average at every **30** years

**5 borehole**: POM-SAT forward model
Interesting questions

- What is the skill of **forcings**?
- What is the skill of each **proxy**?
- How does the **noise** in proxies affect the reconstruction?
  - **Perfect proxies**: proxies directly generated from the local/regional temperatures
  - **Contaminated proxies**: The signal to noise ratio is chosen 1:4 in terms of their variance
Variations of the full model

- Compare reconstructions with forcings and without forcings:

\[(T_1, T_2) = \beta_0 + \epsilon_5, \quad \epsilon_5 \sim \text{AR}(2)(\sigma^2_5, \phi_{13}, \phi_{23})\]

- Oracle experiment:
Reconstruction using all the original local temperatures rather than their transformed proxies as a baseline to evaluate reconstruction performance:

\[T_{l_i} | (T_1, T_2) = \beta_{0_i} + \beta_{1_i}(T_1, T_2) + \epsilon_i, \quad \epsilon_i \sim \text{AR}(2)(\sigma^2, \phi_1, \phi_2),\]
Experiments

(a) Perfect proxies With forcings
(b) Perfect proxies Without forcings
(c) Contaminated proxies With forcings
(d) Contaminated proxies Without forcings

For each (a), (b), (c) and (d):

(i) local/regional temperature series (PF)
(ii) Tree ring only (R)
(iii) Tree ring + Pollen (RP)
(iv) Tree ring + Borehole (RB)
(v) Tree ring + Borehole + Pollen (RBP)
The figure shows two types of plots, one with bias on the y-axis and the other with RMSE of the reconstruction on the y-axis. Each plot is divided into four quadrants, corresponding to different combinations of forcings and proxies.

- **Bias Plot:**
  - The x-axis represents models (PF, R, RP, RB, RBP).
  - The y-axis represents bias, ranging from -0.2 to 0.4.
  - The quadrants are labeled as follows:
    - Perfect proxies with forcings
    - Perfect proxies without forcings
    - Contaminated proxies with forcings
    - Contaminated proxies without forcings
  - Crosses indicate the level of smoothing (year) for each model.

- **RMSE Plot:**
  - The x-axis represents level of smoothing (year) from 0 to 100.
  - The y-axis represents RMSE of the reconstruction, ranging from 0.1 to 0.4.
  - The quadrants are labeled as follows:
    - Perfect proxies with forcings
    - Perfect proxies without forcings
    - Contaminated proxies with forcings
    - Contaminated proxies without forcings
  - Dashed lines indicate the models (PF, R, RP, RB, RBP) with their respective RMSE values.
Contaminated proxies With Forcings

![Graph showing temperature targets and various forcings.](image_url)
Discussion

- **Forcings** play an important role in the temperature reconstruction.
- **Pollen and borehole** together can greatly improve the **calibration** of the reconstruction especially when forcings are not available.
- **BHM** provides a **flexible** framework to integrate information from different sources.
- The results are based on the climate from **one model run**, so they may not replicate very well given a different model run.
- **Borehole model** needs further investigation.