USC University of Southern California

Probabilistic Models of Past Climate Change

Julien Emile-Geay

USC College, Department of Earth Sciences

SIAM International Conference on Data Mining Anaheim, California April 27, 2012

Probabilistic Models of Past Climate Change

Julien Emile-Geay (USC)

with: **Dominique Guillot (USC)**
Bala Rajaratnam (Stanford)
Tapio Schneider (CalTech)

acknowledgements: **Jianghao Wang** (USC)

Outline:

1. Mathematical Problem
2. Gaussian Graphical Models
3. Some results
4. Error estimates

CLIMATE DYNAMICS @ USC

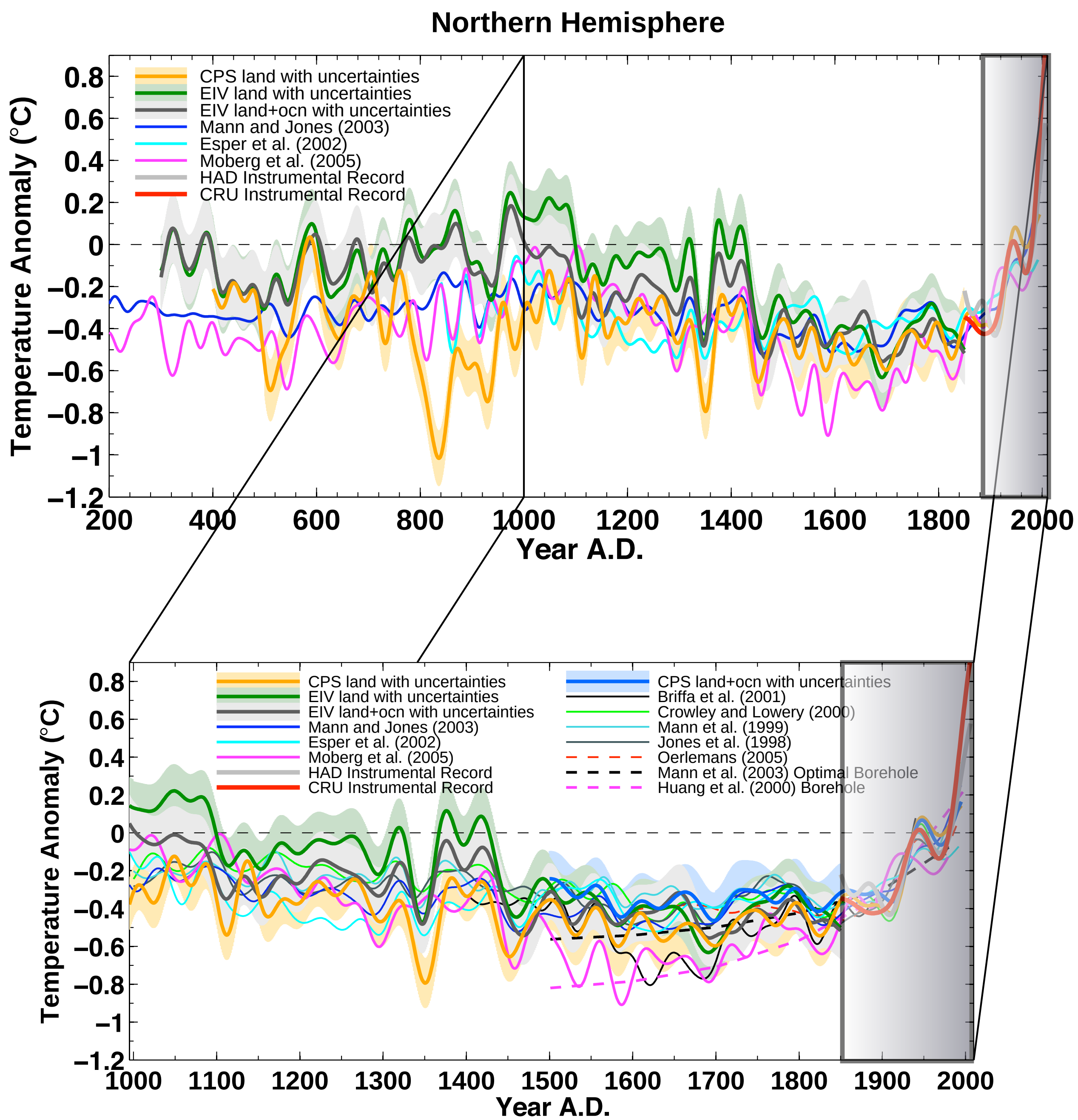
Reconstructing Past Climates

Why paleoclimatology?

- Is it warmer now than in AD 1000? ("*Hockey Stick" problem*)
- Is the rate of warming anomalous?
- What are the spatiotemporal characteristics of natural climate variability?
- How (un)certain is all this?

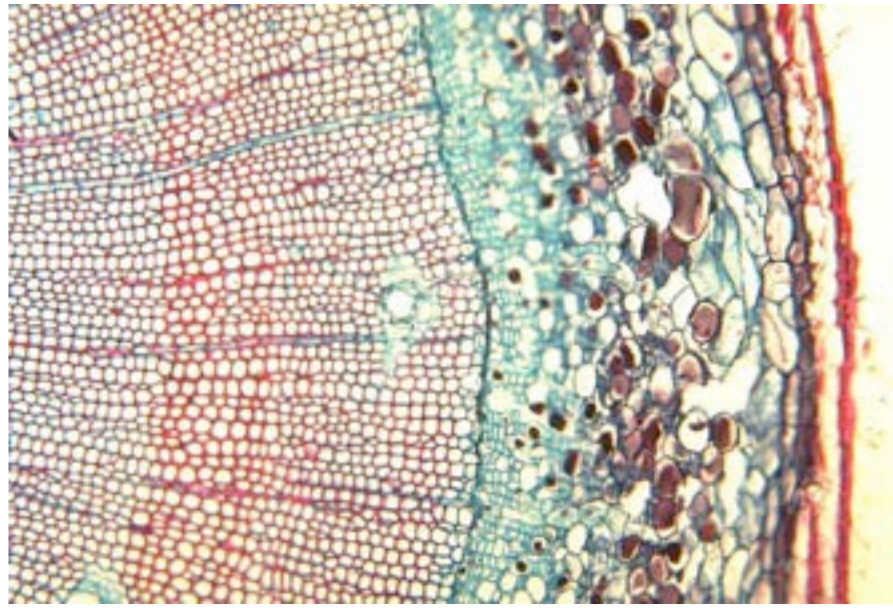
Statistical challenges

- Short training set (calibration)
- Very high-dimensional ($p > n$)
- Noisy, autocorrelated predictors
- No straightforward spatial model



Mann et al., PNAS, 2008

High-resolution paleoclimate proxies



Tree Rings

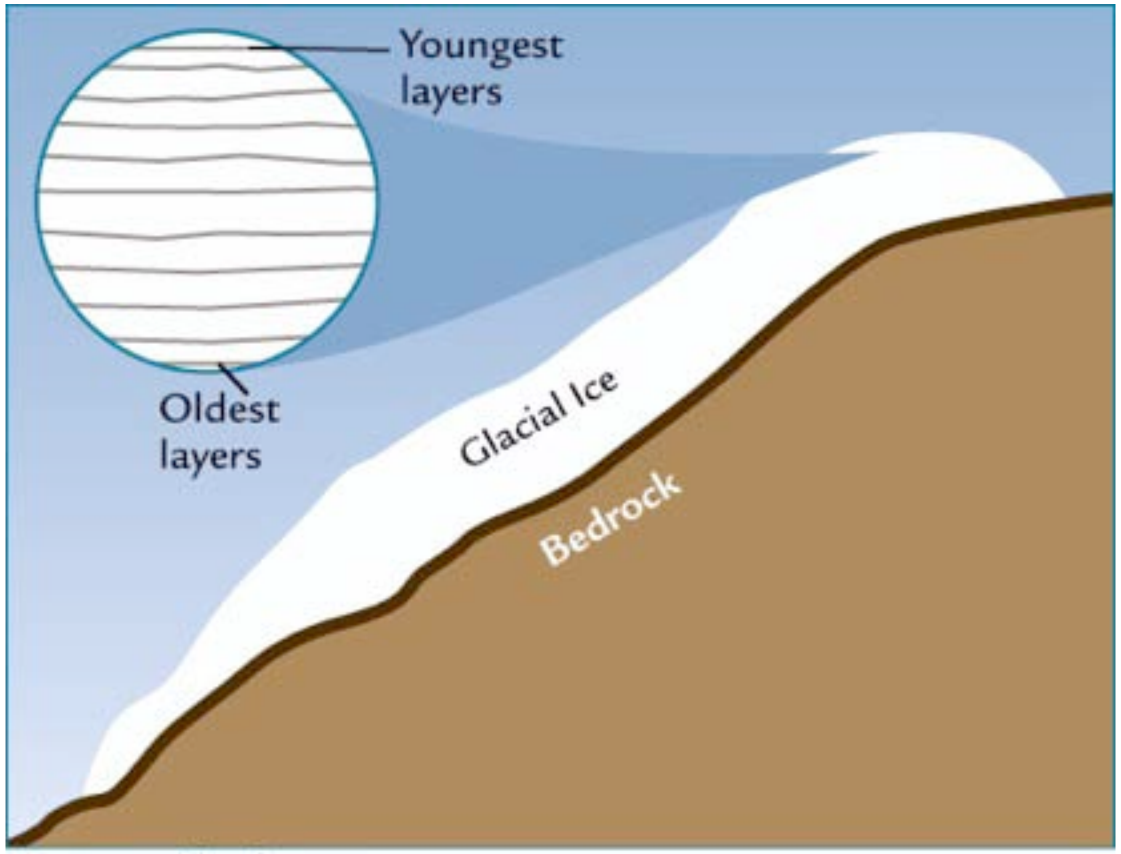
Speleothems

Ice cores

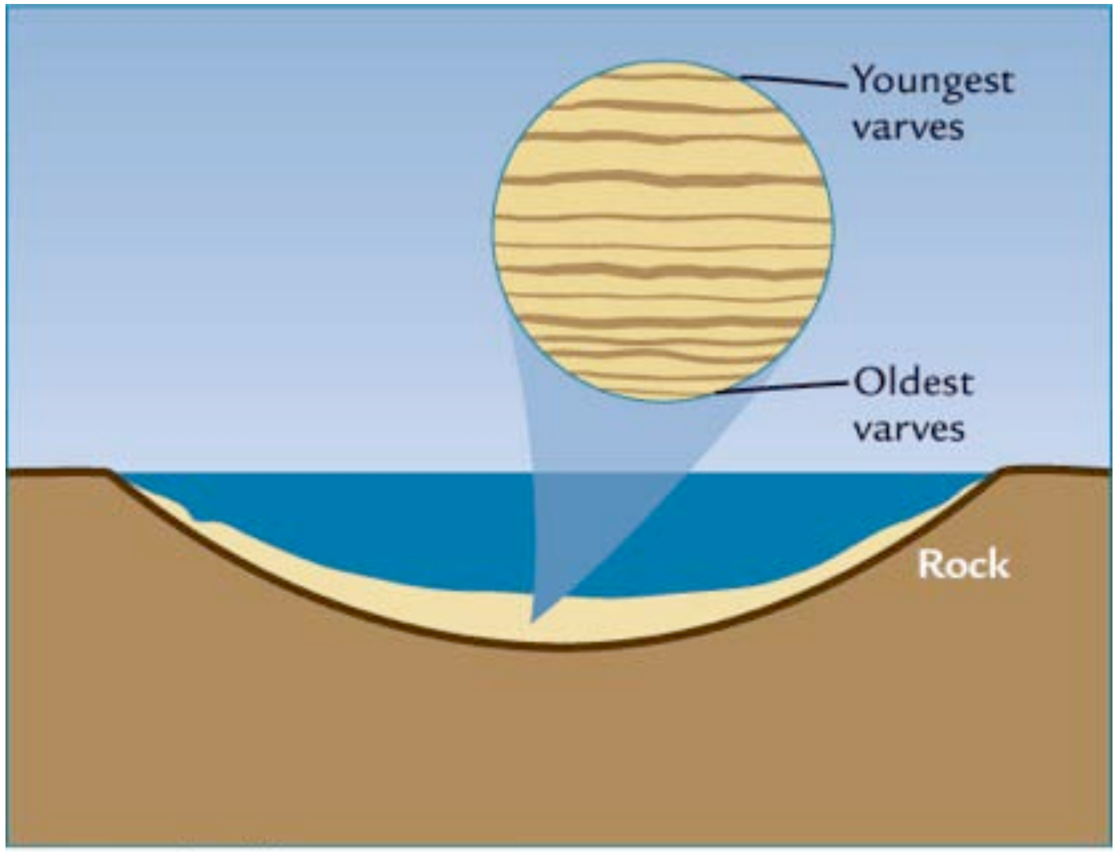
Sediment cores

Corals

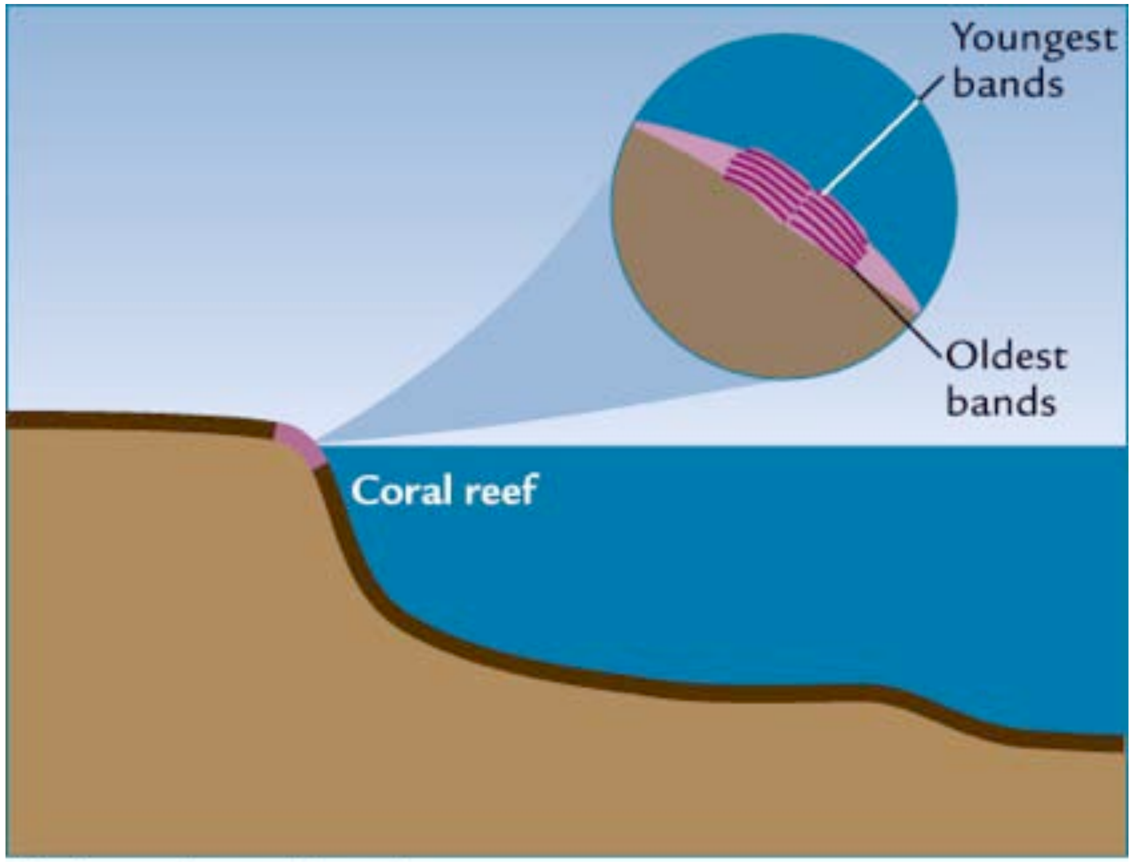
High-precision dating



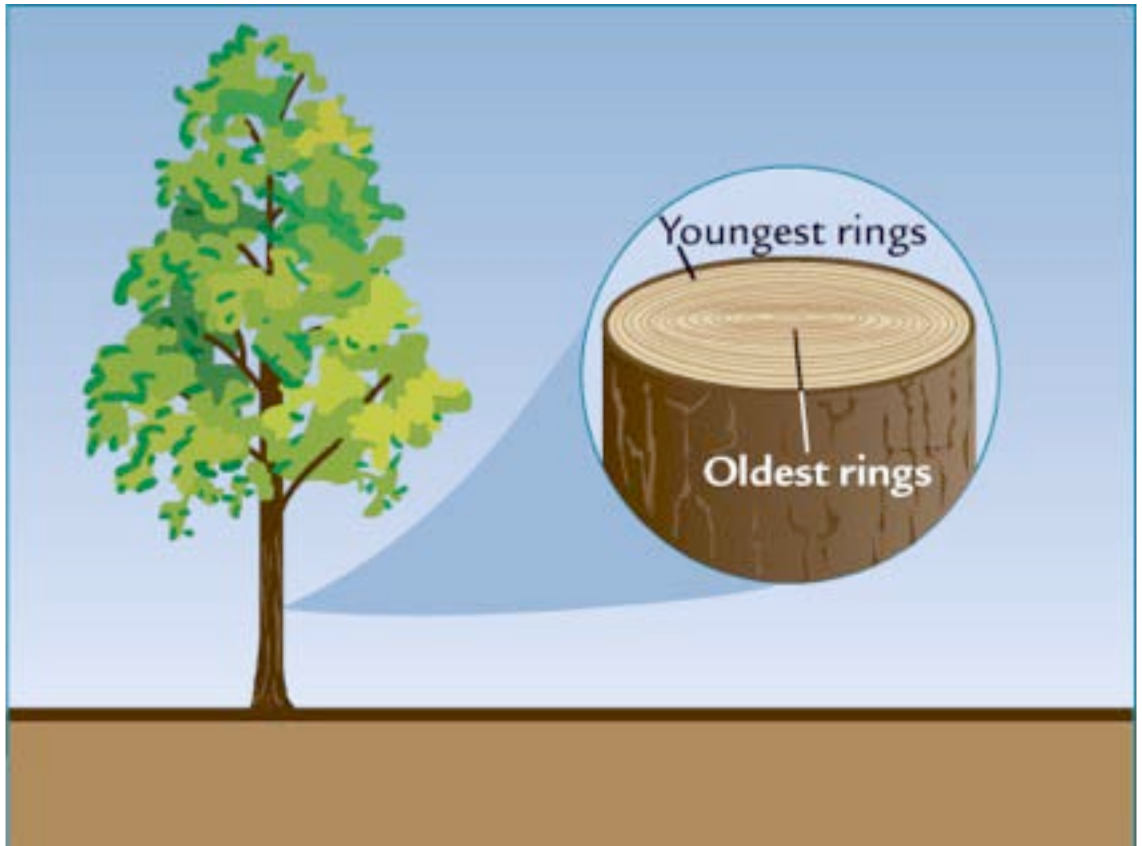
A Annual ice layers



B Annual sediment varves



D Annual coral bands



C Annual tree rings

Ruddiman, 2006

Climate Field Reconstruction

A missing data problem

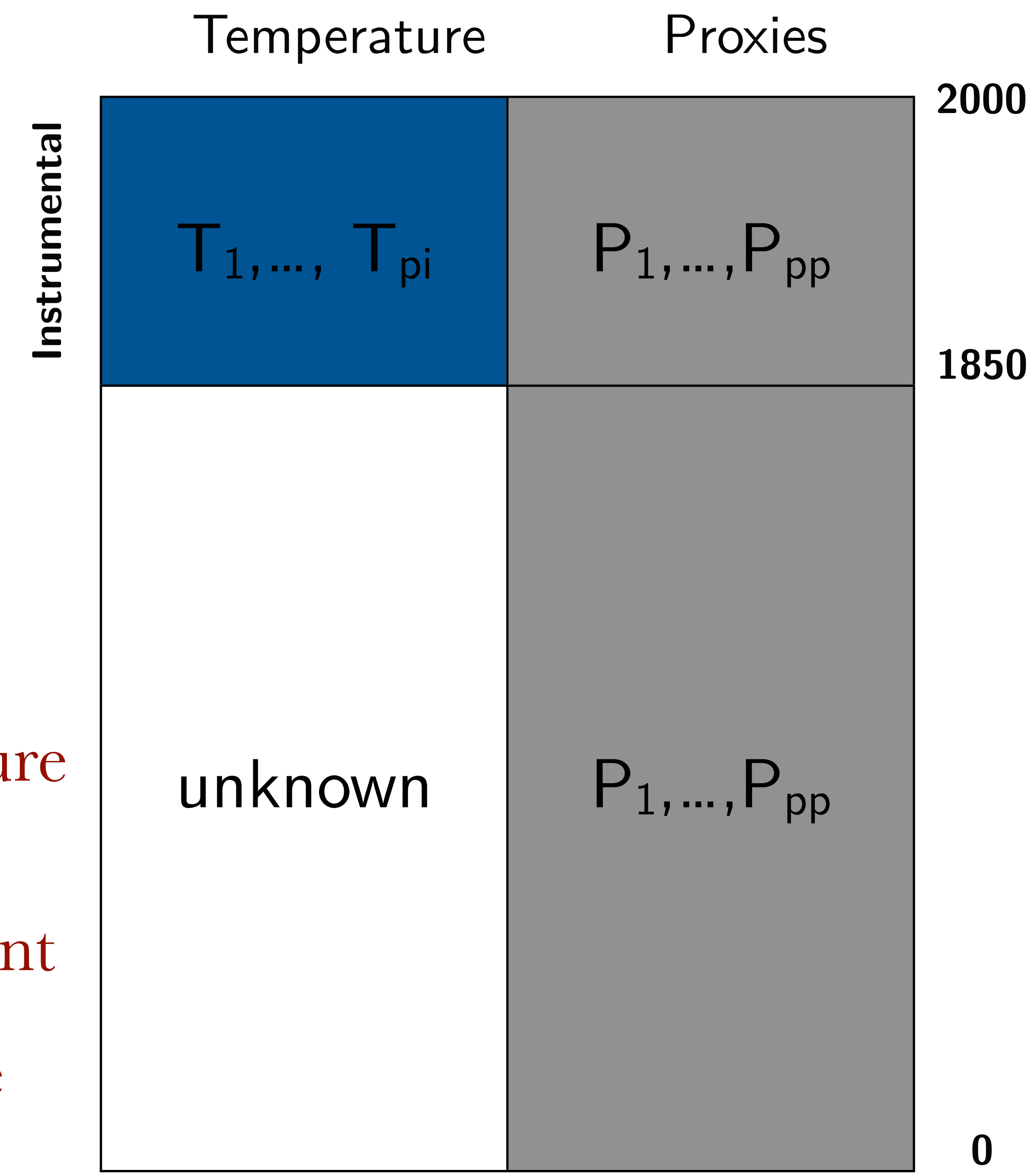
- backcast T from proxy observations
- multivariate inference

A high-dimensional problem

- e.g. *Mann et al* [2008] database
- $p = p_i + p_p = 1732 + 1138 \gg n = 150$

Covariance matrix

- captures relationship between temperature and proxies
- sample covariance matrix is rank-deficient
- estimation is impossible from the sample covariance matrix: it must be *regularized*



Regularized Covariance Estimation

(1) Maximum Likelihood Estimation using ℓ_2-penalized likelihood and Expectation-Maximization [*Dempster, Laird and Rubin*, 1977]

$$L_2(\Sigma, h) = \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S) - \frac{n}{2}|\Sigma^{-1}| + h\sum_{i<j} ||\sigma^{ij}||_2^2$$

(2) Regularized Solution:

$$\hat{\Sigma}_{aa} = \mathsf{V}\boldsymbol{\Lambda}^2\mathsf{V}^{\mathrm{T}} \qquad \mathsf{F} \equiv \boldsymbol{\Lambda}^{\dagger}\mathsf{V}^{\mathrm{T}}\hat{\Sigma}_{am}$$

(Fourier coefficients)

$$\hat{\mathsf{B}} = \mathsf{V}\,\mathrm{Diag}(f_j)\,\boldsymbol{\Lambda}^{\dagger}\,\mathsf{F}$$

with f_j the filter factors

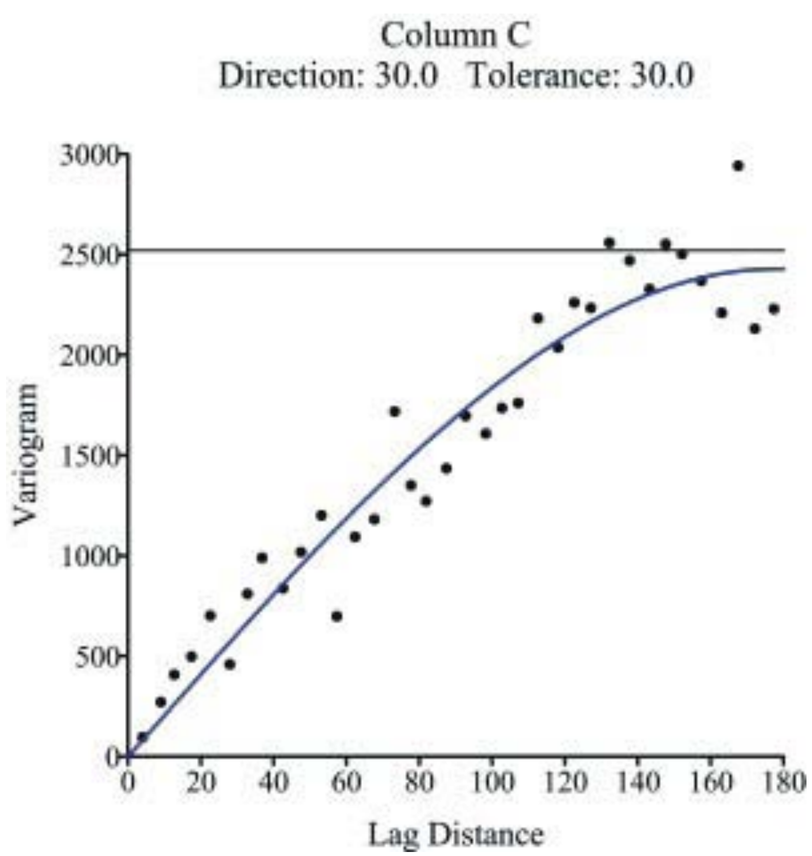
Tikhonov Regularization ("Ridge regression")

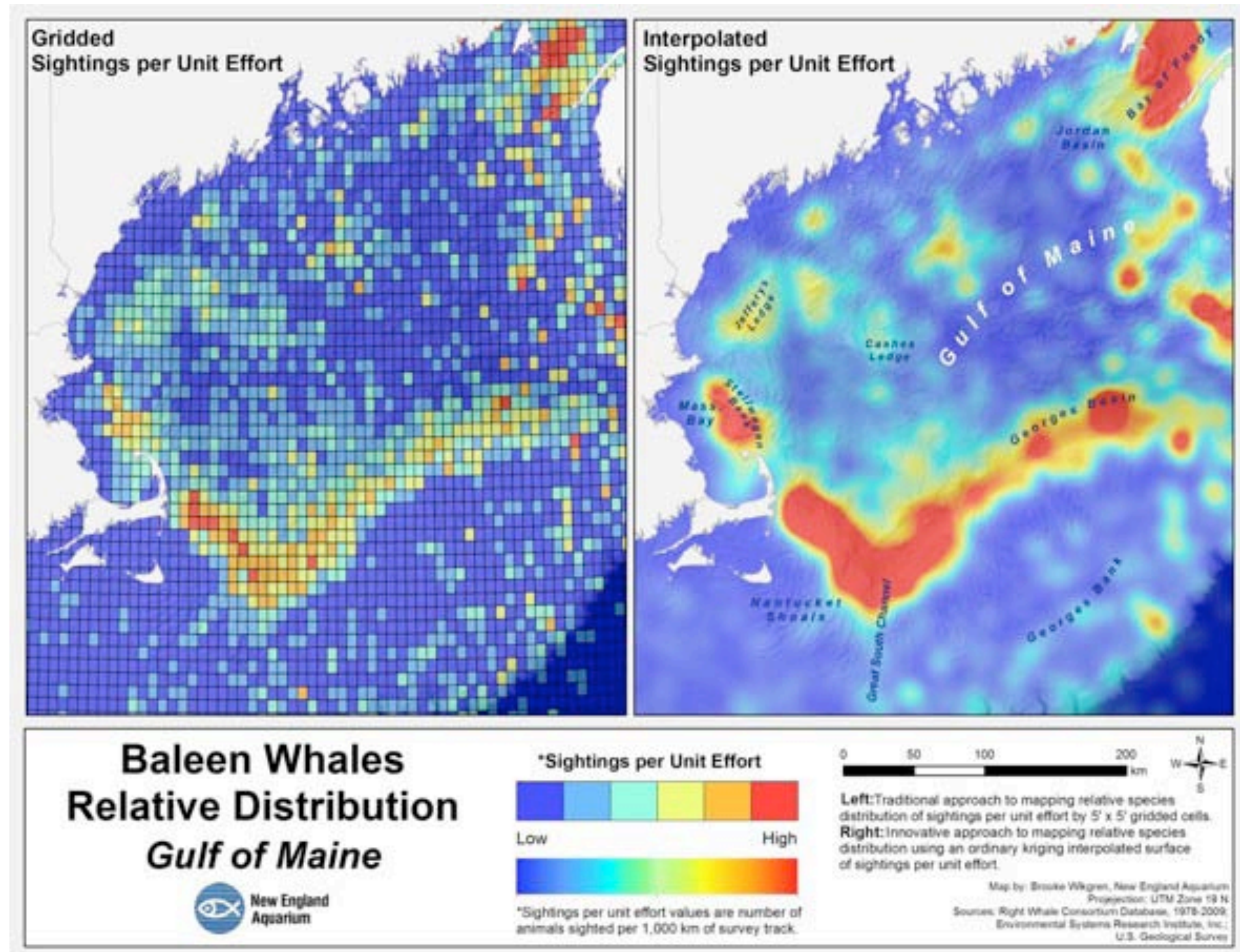
$$f_j = {\lambda_j}^2/({\lambda_j}^2 + h^2)$$

Explicit Spatial Modeling

e.g. “kriging”

$$C_{ij} = f(d_{i-j})$$

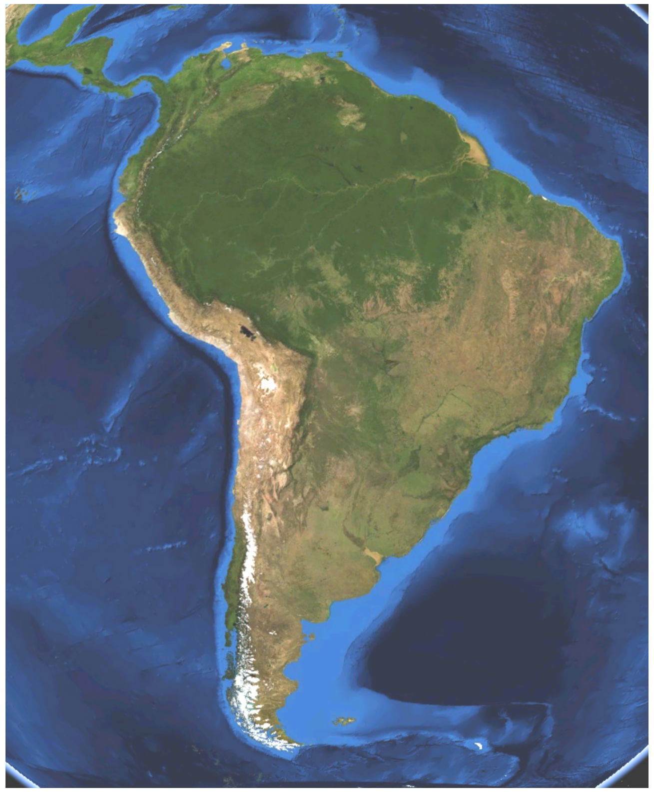




Limitations:

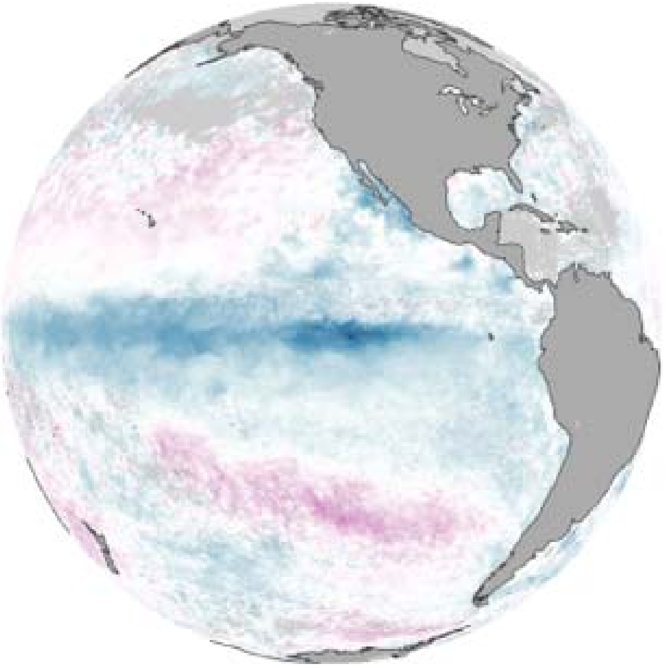
- isotropic
- rigid
- subjective

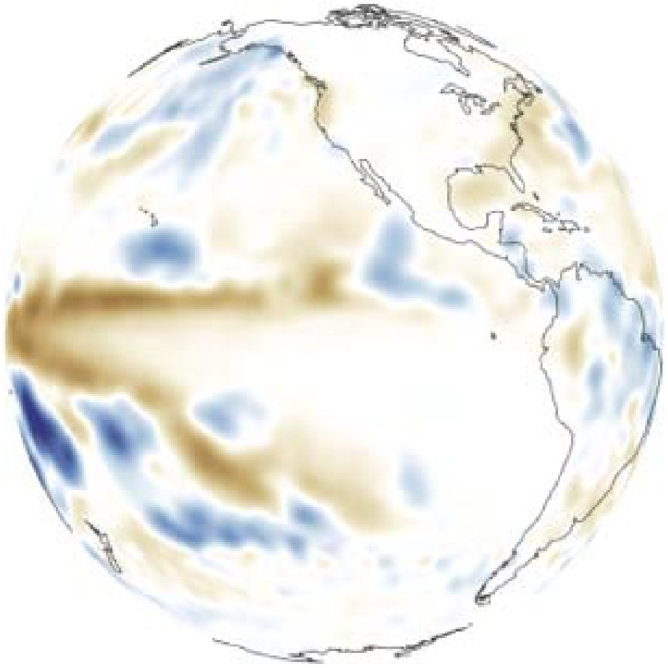
Graphical models: spatial modeling



Land/Ocean boundaries

Mountain ranges





Teleconnections / climate patterns

Marginal vs Conditional Independence

$$\text{Marginal independence: } X_i \perp\!\!\!\perp X_j \Leftrightarrow \Sigma_{ij} = 0$$

$$\text{Conditional independence: } X_i \perp\!\!\!\perp X_j | \{\text{rest of variables}\} \Leftrightarrow \Omega_{ij} = 0.$$

Example:

$$\Omega = \begin{pmatrix} 40.5423 & \mathbf{0} & 0.0048 & 5.6675 & -39.2268 & -5.6599 \\ \mathbf{0} & 2.0969 & 1.5166 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0.0048 & 1.5166 & 2.0969 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 5.6675 & \mathbf{0} & \mathbf{0} & 39.7654 & \mathbf{0} & -39.2357 \\ -39.2268 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 39.7300 & \mathbf{0} \\ -5.6599 & \mathbf{0} & \mathbf{0} & -39.2357 & \mathbf{0} & 39.7177 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1.0000 & 0.0035 & -0.0048 & -0.0759 & 0.9873 & 0.0676 \\ 0.0035 & 1.0000 & -0.7233 & -0.0003 & 0.0034 & 0.0002 \\ -0.0048 & -0.7233 & 1.0000 & 0.0004 & -0.0047 & -0.0003 \\ -0.0759 & -0.0003 & 0.0004 & 1.0000 & -0.0749 & 0.9771 \\ 0.9873 & 0.0034 & -0.0047 & -0.0749 & 1.0000 & 0.0667 \\ 0.0676 & 0.0002 & -0.0003 & 0.9771 & 0.0667 & 1.0000 \end{pmatrix}$$

Discovering conditional independence relations

Exploiting conditional independence relations:

- Conditional independence relations are inherent to climate fields:
- Knowledge of such relations $\Rightarrow$ zeros in Ω $\Rightarrow$ **dimension reduction**;
- Can be discovered using ℓ_1 **type** optimization methods.

Graphical lasso: *Friedman, Hastie & Tibshirani [2008]*

$$\max_{\Omega > 0} \ \text{log-likelihood}(\Omega) + \rho \underbrace{\sum_{i,j} |\Omega_{ij}|}_{\|\Omega\|_1}.$$

Graphical maximum likelihood estimate:

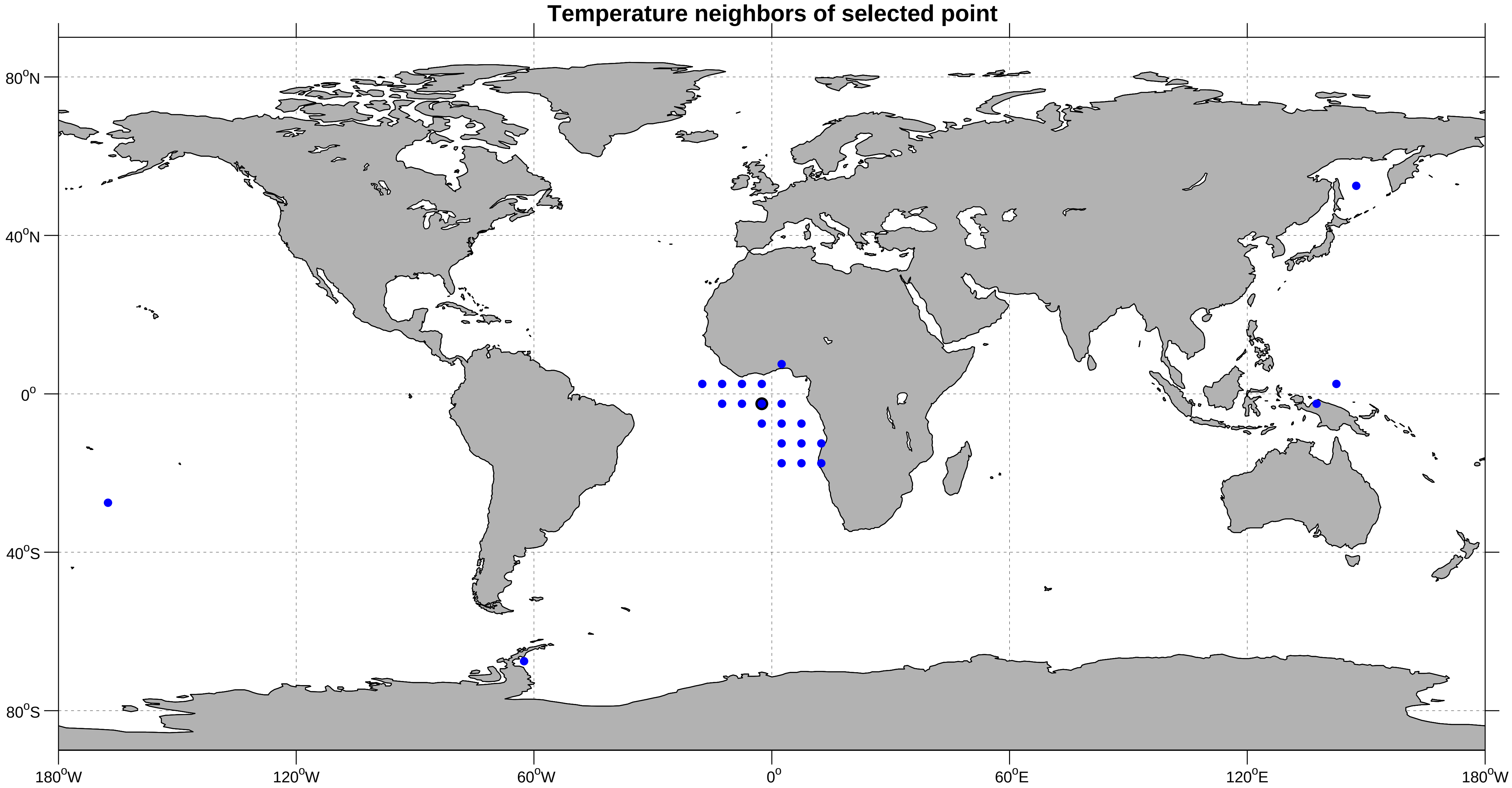
$$\hat{\Sigma}_G = \max_{\Sigma^{-1} \in \mathbb{P}_G^+} \ \text{likelihood}(\Sigma)$$

"Best" covariance matrix compatible with the CI structure.

Benefits:

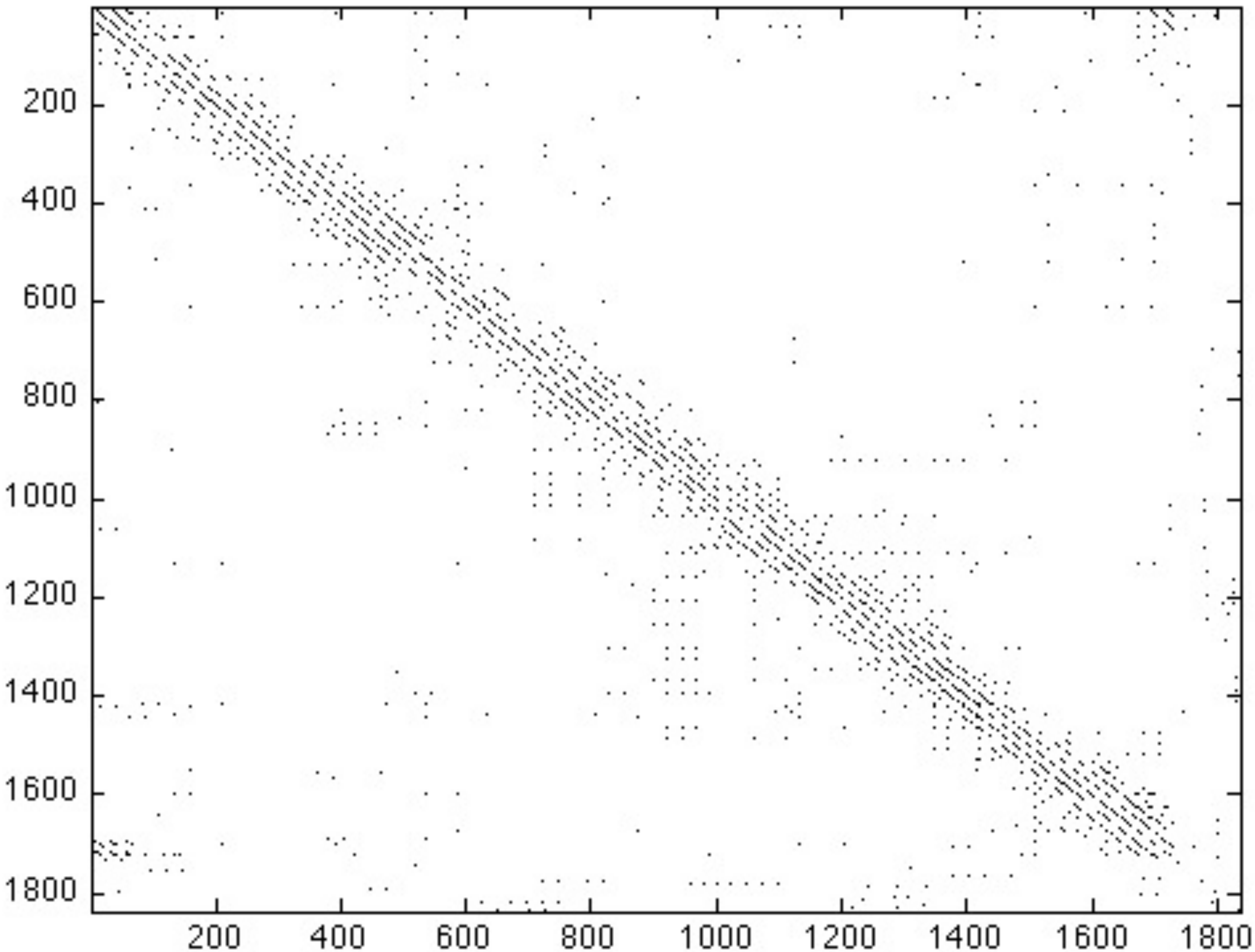
- Adding a ℓ_1 penalty favors **sparse** estimates of Ω;
- Sparse estimates achieve the necessary **dimension reduction** for proper estimation of Σ.

Flexible covariance representation



HadCRUT3v data, 1850-2000

Example of graph (HadCRUT3v data)



The Graphical EM (GraphEM) algorithm

Dempster, Laird & Rubin, 1977

Schneider, 2001

(1)

$$\left\{ \begin{array}{c} \mu_0, \Sigma_0 \\ \text{CI structure} \end{array} \right\}$$

(2)

Compute

$$\hat{\Sigma} = \hat{\Sigma}_G$$

(Regularization)

Compute regression coefficients

$$\hat{\mathsf{B}} = \hat{\boldsymbol{\Sigma}}_{aa}^{+} \hat{\boldsymbol{\Sigma}}_{am}$$

(3)

$$\hat{\mathsf{x}}_m = \hat{\boldsymbol{\mu}}_m + (\mathsf{x}_a - \hat{\boldsymbol{\mu}}_a)\hat{\mathsf{B}}$$

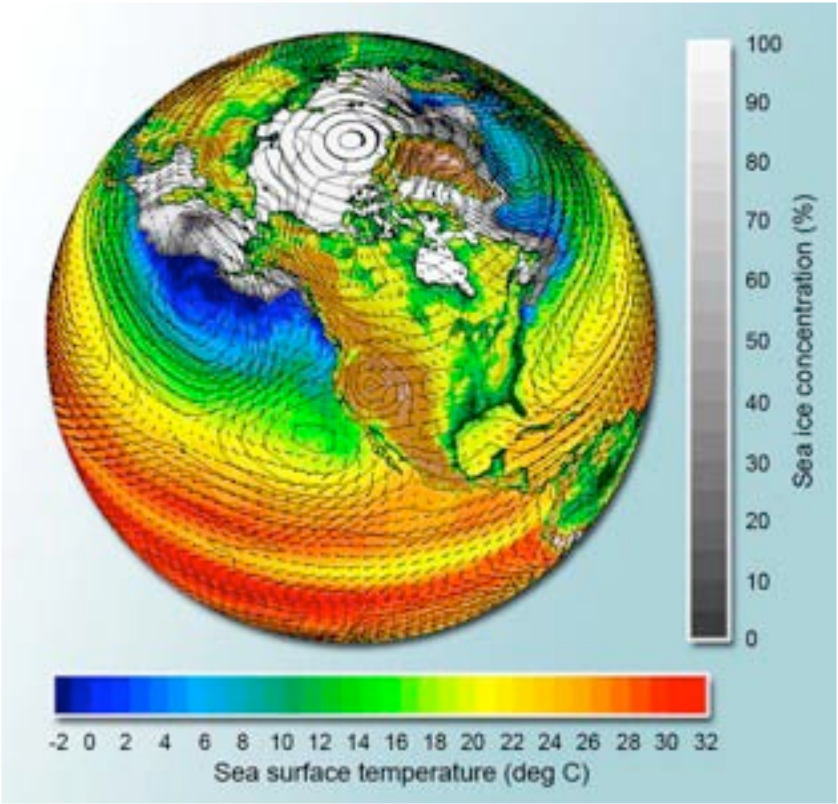
Estimate unknown values from the available ones

(4)

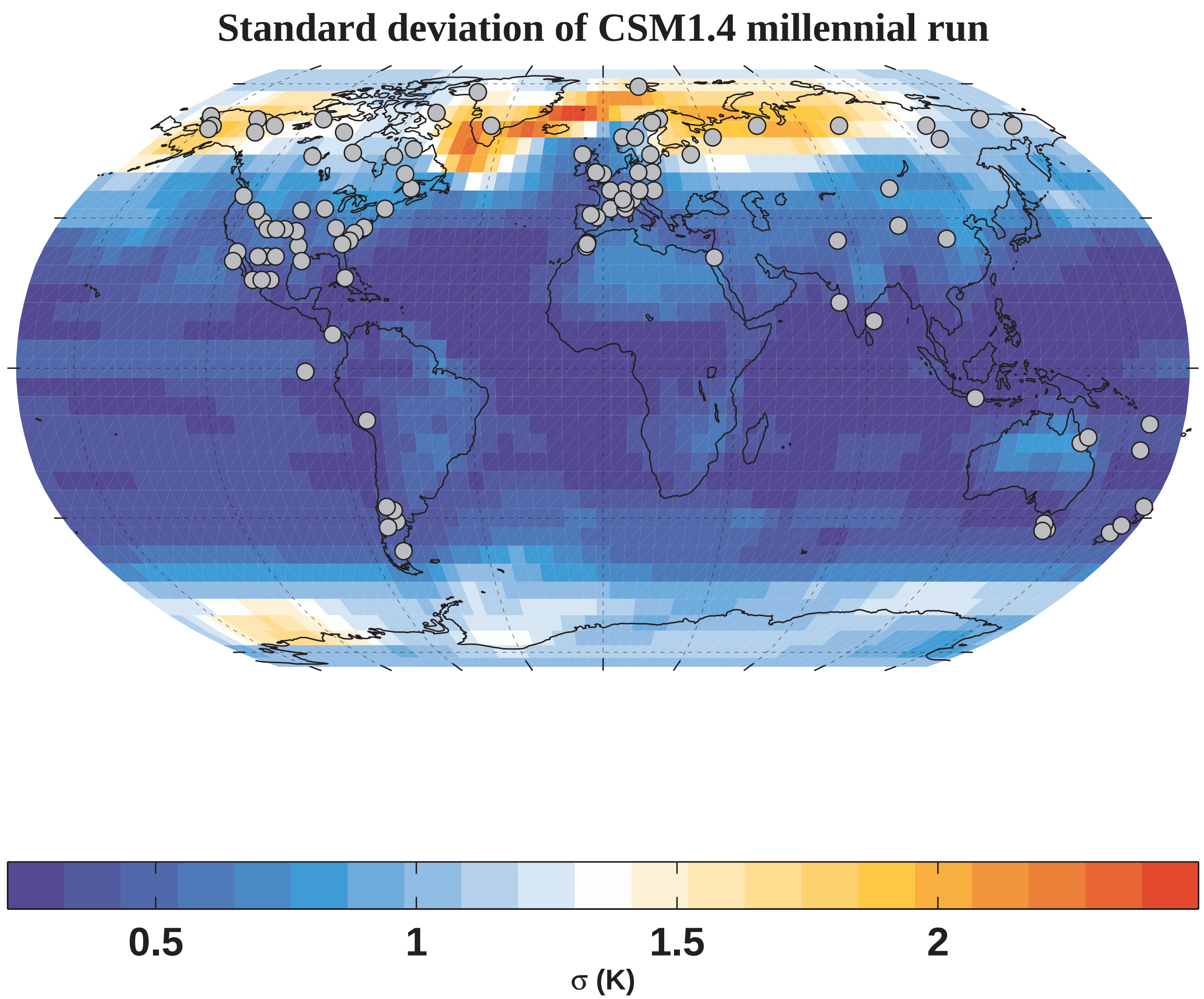
Update

(5) $\{\mu, \Sigma\}$

Guillot et al, JASA, submitted



A virtual climate laboratory



CSM1.4 specs:

- Coupled General Circulation Model
- Plausible "surrogate climate"
- Generate pseudoproxies as statistically-degraded, subsampled version of the temperature field

Pseudoproxy Tests

The GraphEM methodology is tested on <u>synthetic proxies</u> derived from a forced simulation of the NCAR CSM1.4 model (including volcanic and solar forcing)

$$P_p(s,t) = T(s,t) + \xi(s,t)/\mathrm{SNR}$$

where ξ is a standard, uncorrelated Gaussian process

Signal to noise ratio:

$$\mathrm{SNR} = \frac{\rho}{\sqrt{1-\rho^2}}$$

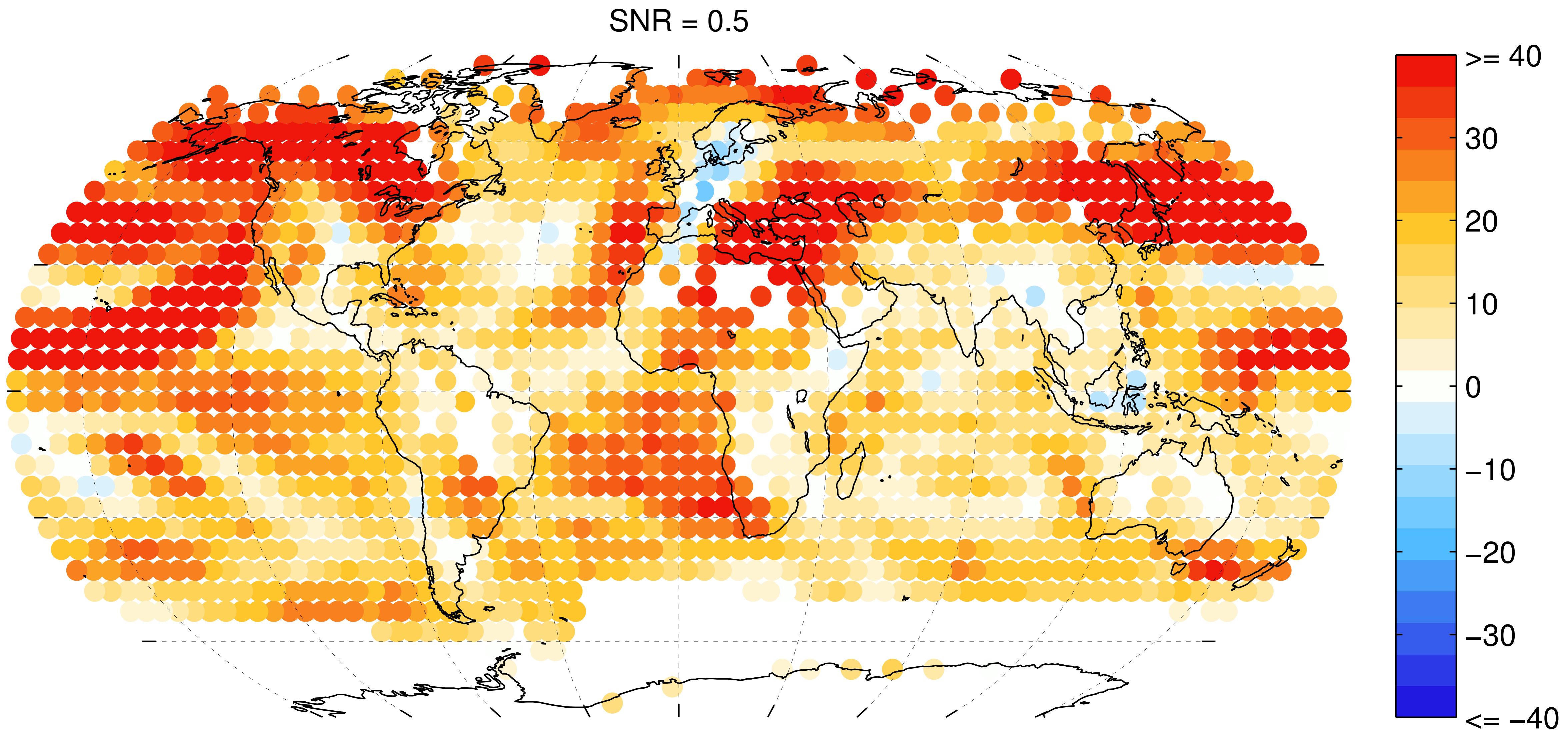
Test Statistic:

$$\mathrm{MSE} = \sum_i (\hat{y}_i - y_i)^2$$

Case	SNR = ∞	SNR = 1	SNR = 0.5	SNR = 0.25

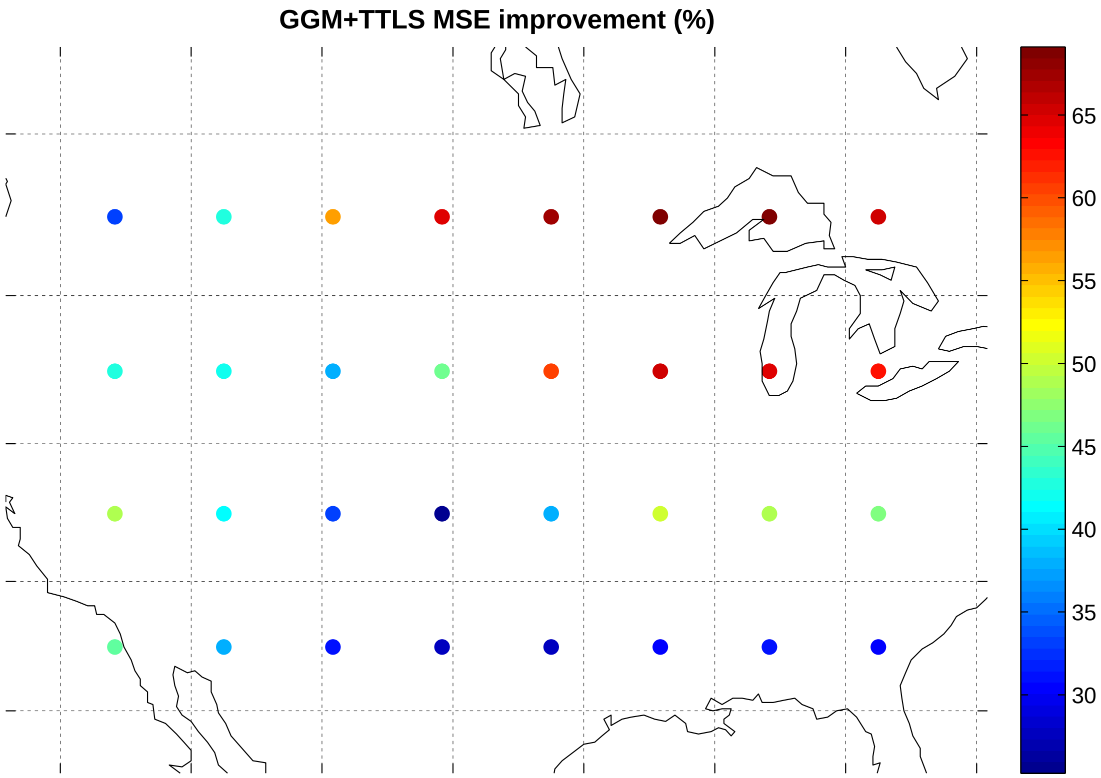
Error reduction

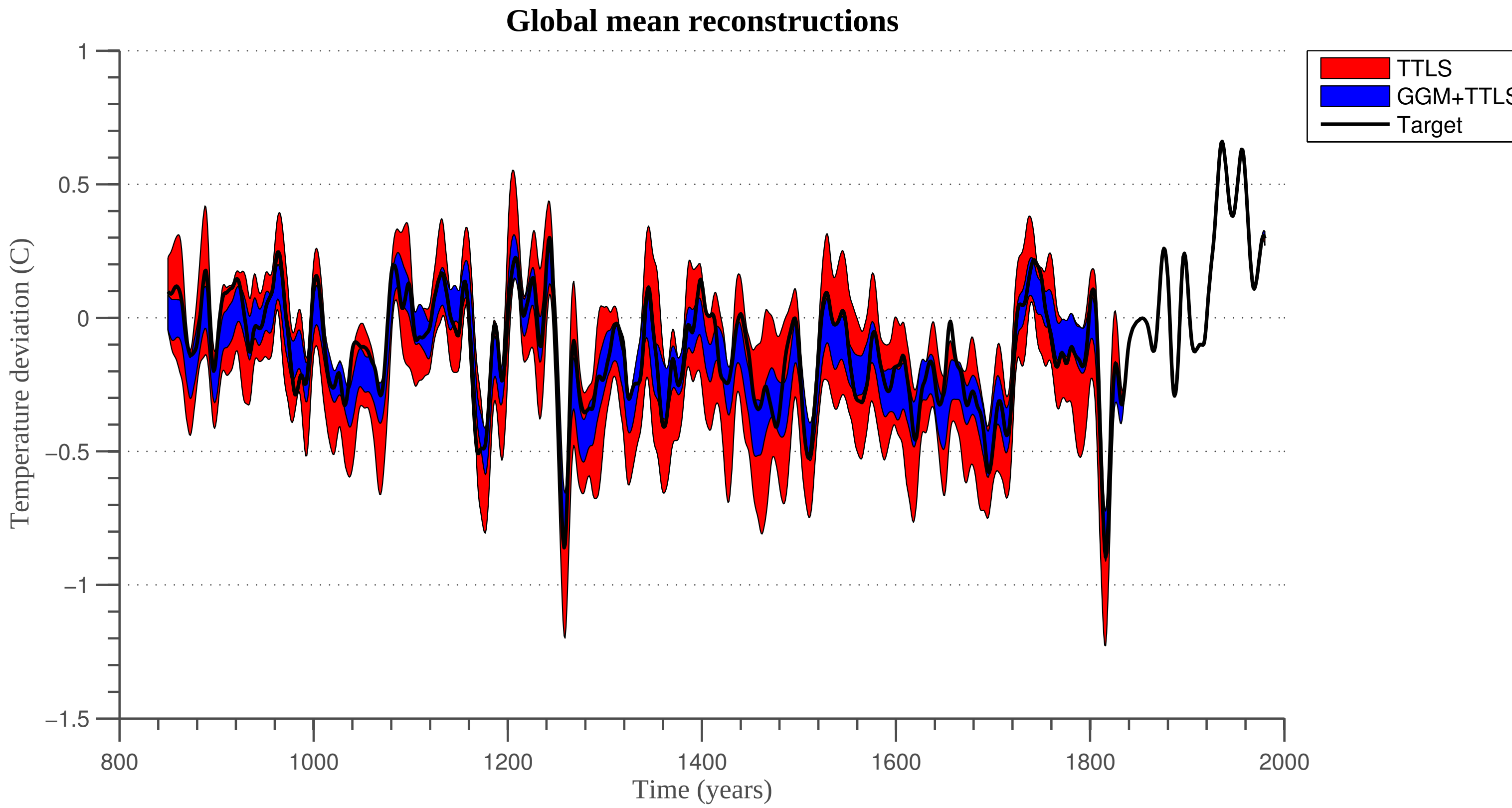
% MSE reduction, GraphEM - RegEM TTLS



North American reconstructions

(100 noise realizations)



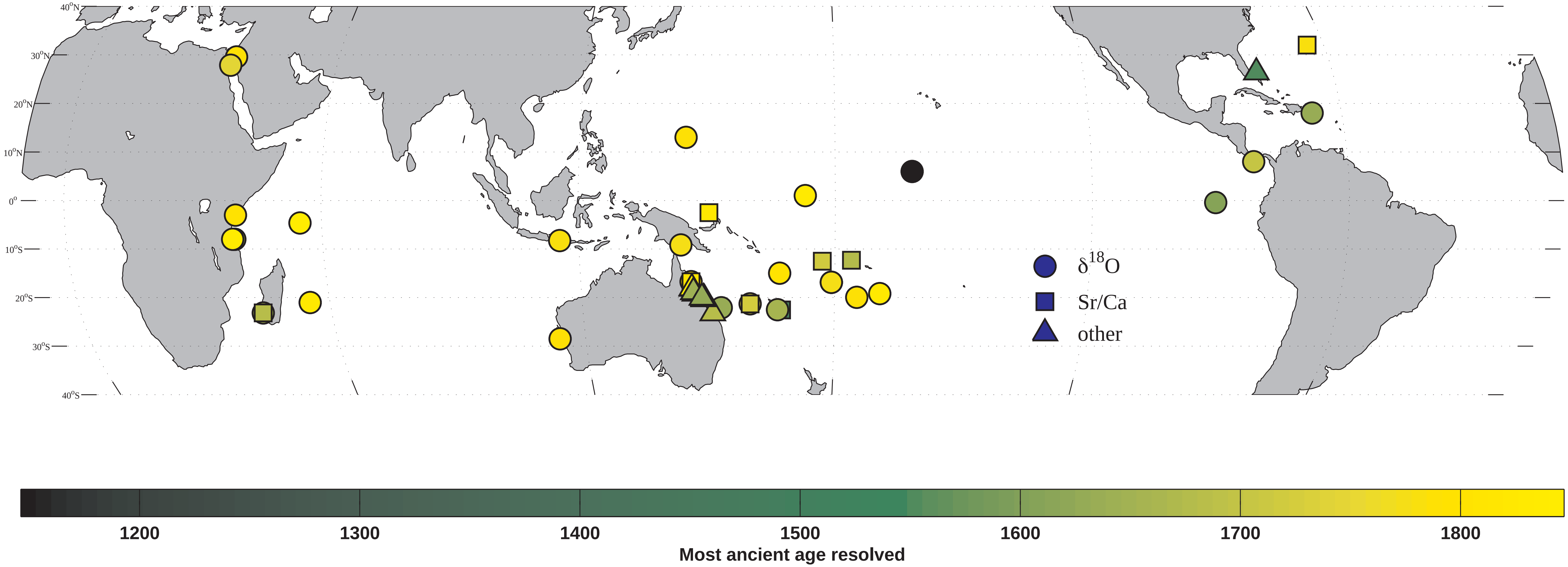


Substantial reduction in MSE

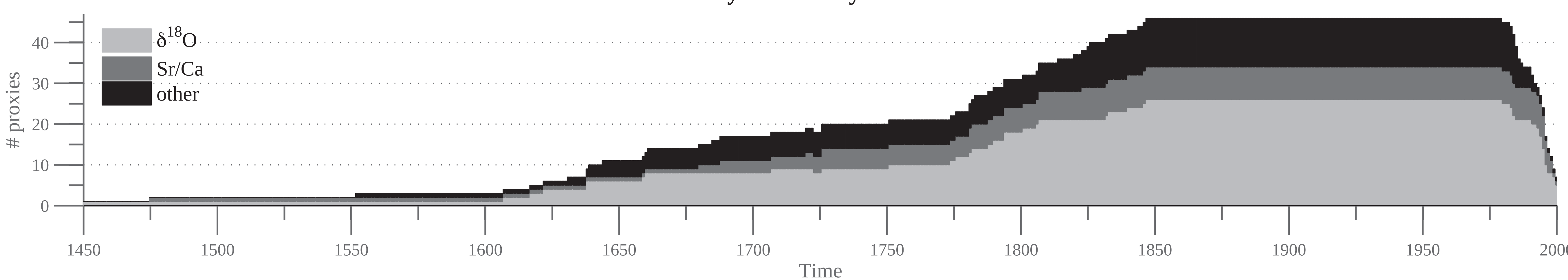
Much improved risk properties

Coral-based sea-surface temperature reconstructions

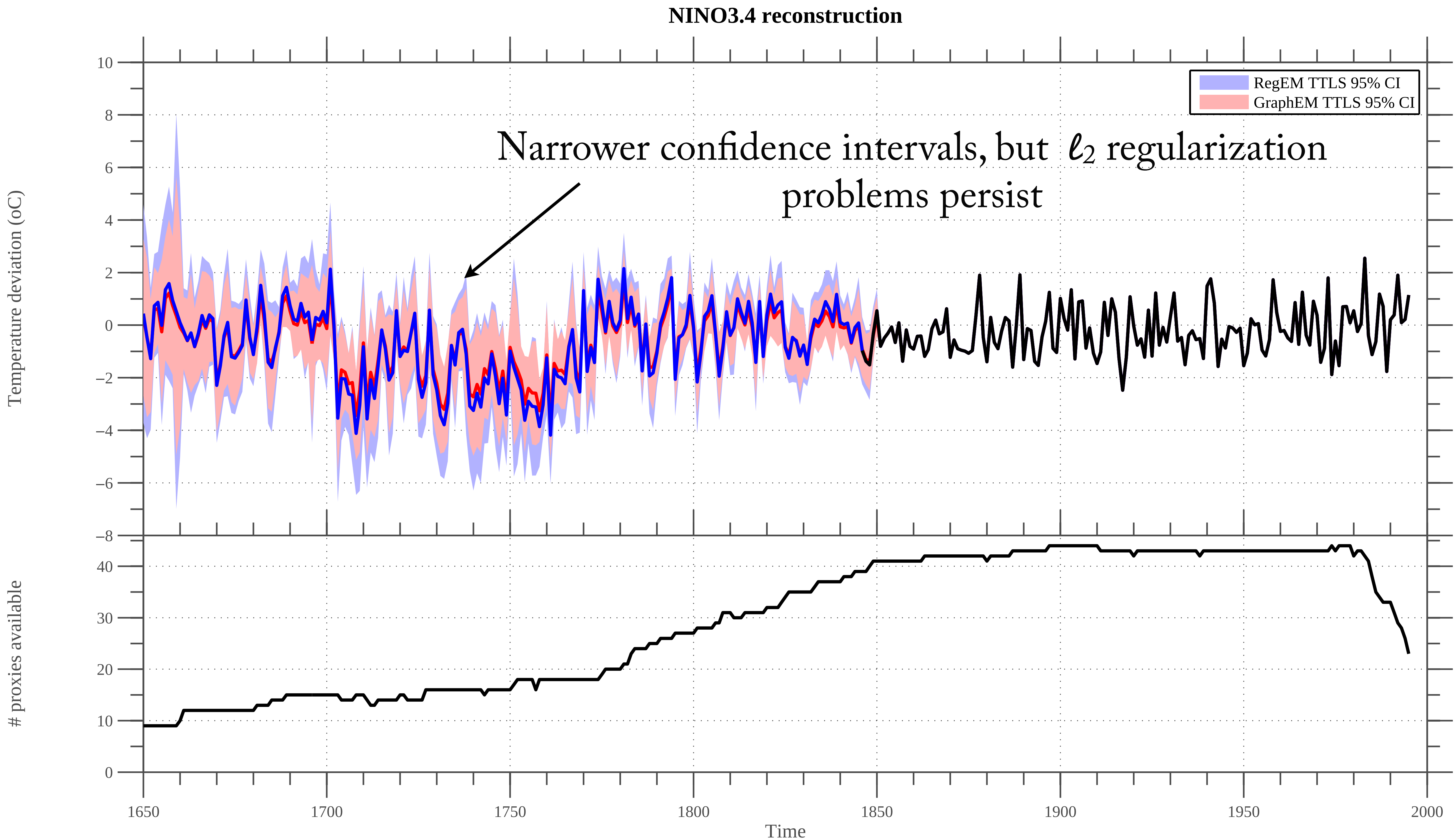
Proxy representation by age (47 records)



Proxy availability over time

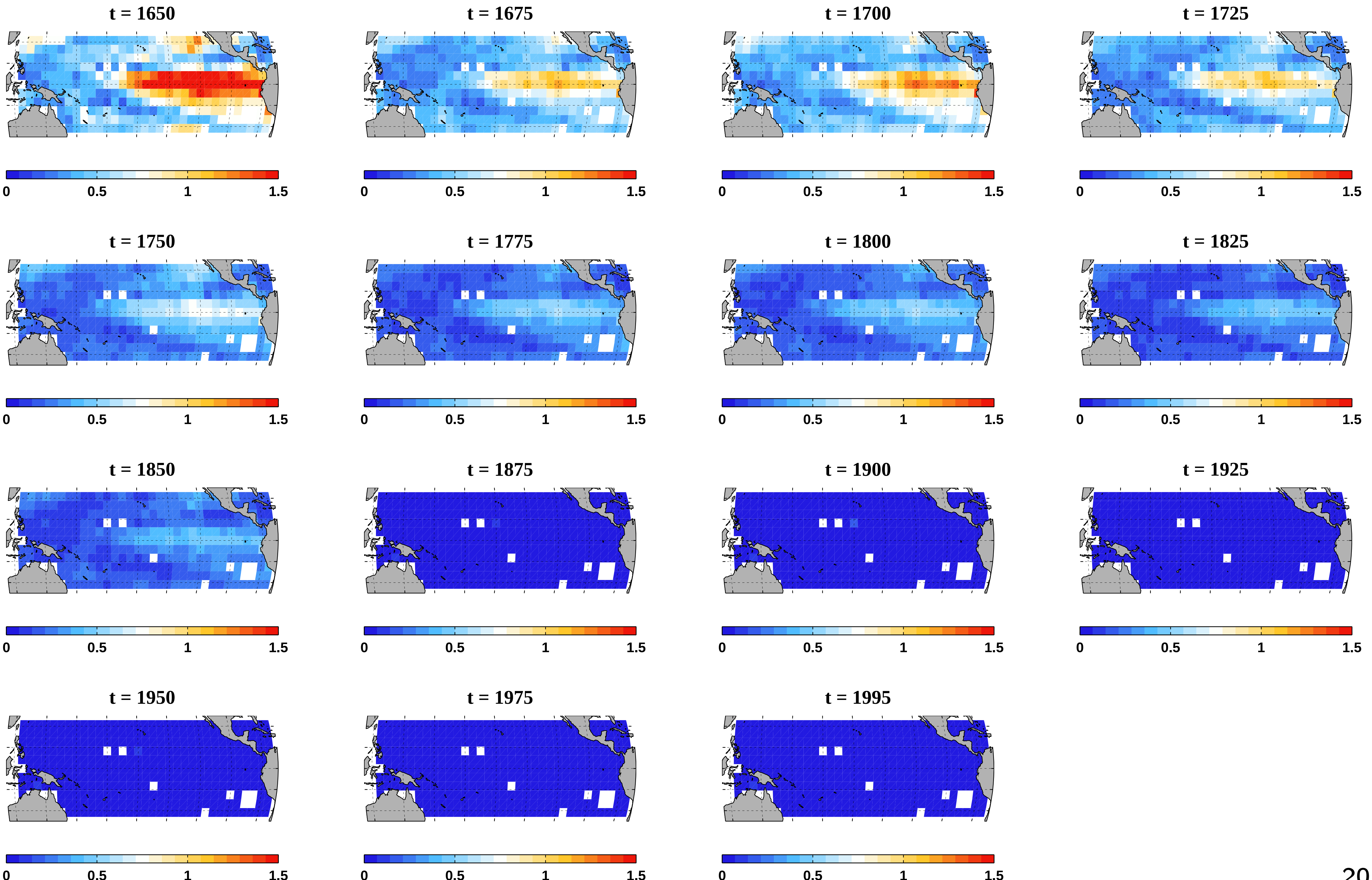


Bootstrap error estimates



Coral-based SST uncertainties: GraphEM TTLS

Spatial estimate of the uncertainties



Conclusions

- **Paleoclimate Reconstructions**
  - High-dimensional, multivariate inference problem
  - Should benefit from latest advances in statistics
- **Gaussian Graphical Models**
  - Enable flexible covariance estimation, reduce errors
  - Model selection, not regularization (ℓ_2 still needed)
  - Choice of graph: spatial truncation?
- **Current and future developments**
  - GraphEM:
    - Analysis of uncertainties (bootstrap interval coverage rate)
    - Application to up-to-date proxy databases (coral, multiproxy)
  - Bayesian Modeling
    - Closed-form Bayes estimators with graphical covariance structure
    - Incorporation into Bayesian hierarchical models

questions, comments, data, preprints:
julieneg@usc.edu

Bayesian Hierarchical Models

Bayes' Theorem

$$\Pr(\text{par}|\text{obs}) = \Pr(\text{obs}|\text{par})\,\Pr(\text{par})$$

Posterior — Likelihood — Prior

3 levels of conditioning:

Y : Climate Process (s,t)

W : Latent Process (s,t) (unobserved)

Z : Observed Process (s,t)

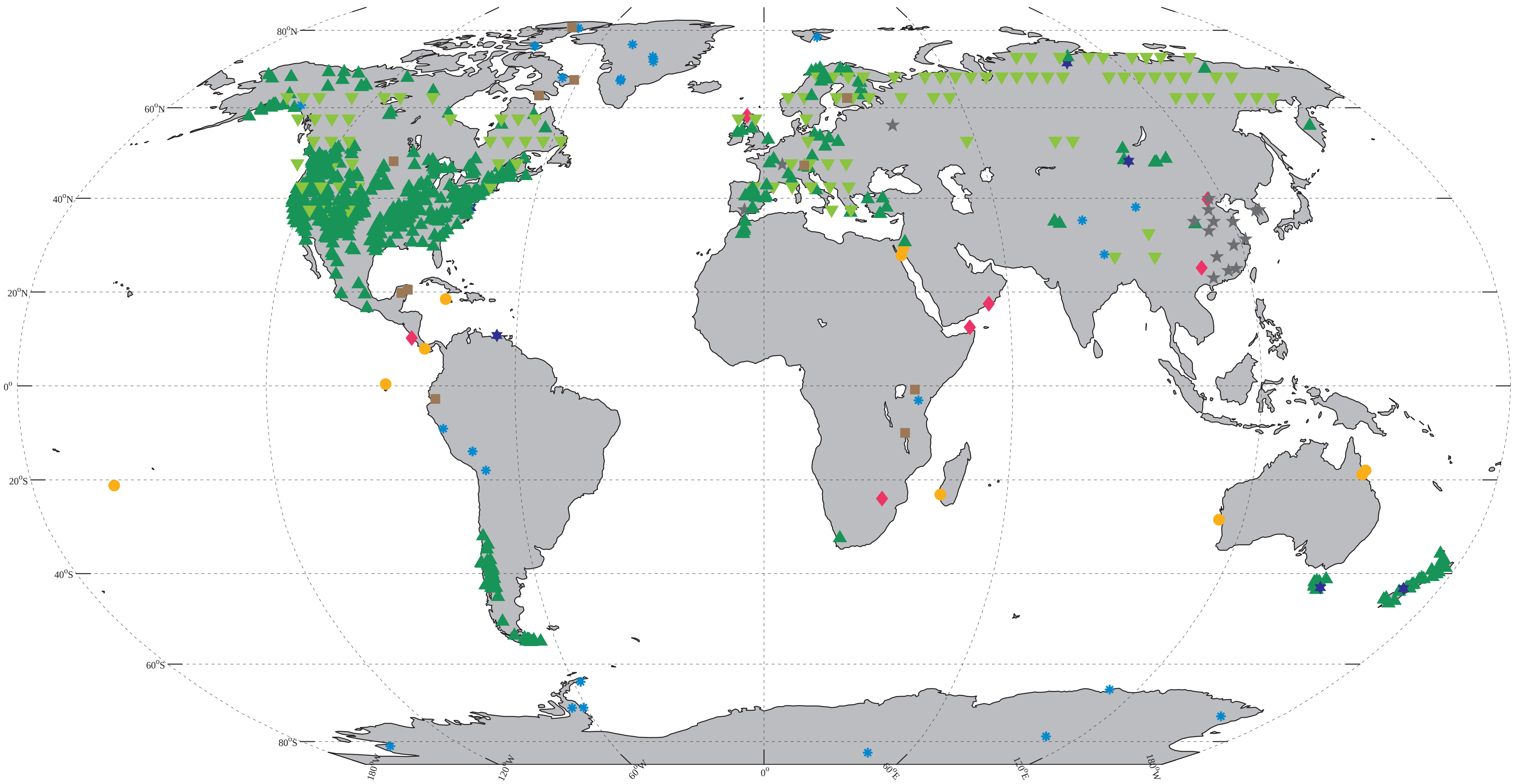
Scientific understanding can be encoded at the appropriate level

$$\boxed{\pi(\mathbf{Y}, \{\mathbf{W}_{I,j}\}, \{\mathbf{W}_{P,j}\}, \boldsymbol{\theta} | \{\mathbf{Z}_{I,j}\}, \{\mathbf{Z}_{P,k}\})} \propto \boxed{f(\mathbf{Y}|\boldsymbol{\theta}) g(\{\mathbf{W}_{I,j}\}, \{\mathbf{W}_{P,k}\}|\mathbf{Y}, \boldsymbol{\theta}) \left[\prod_{j=1}^{N_I} h_{I,j}(\mathbf{Z}_{I,j}|\mathbf{W}_{I,j}, \boldsymbol{\theta})\right] \left[\prod_{k=1}^{N_P} h_{P,k}(\mathbf{Z}_{P,k}|\mathbf{W}_{P,k}, \boldsymbol{\theta})\right]} \boxed{\pi(\boldsymbol{\theta})}.$$

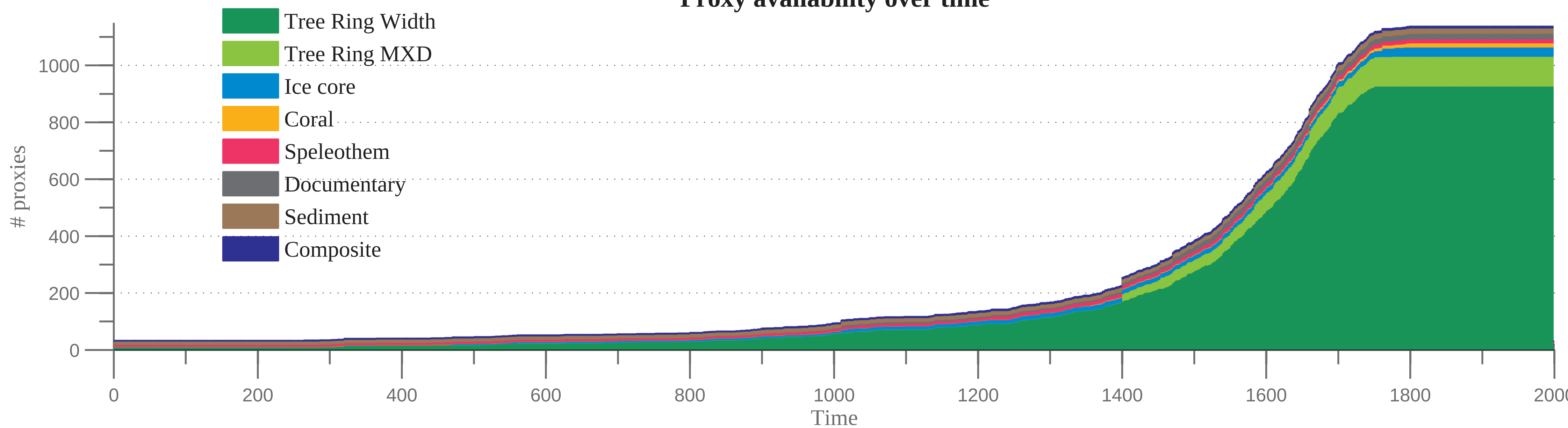
Posterior — Likelihood — Prior

Tingley et al, 2012

Full Proxy Network for Climate Field Reconstruction (1138 records)



Proxy availability over time



CLIMATE DYNAMICS @ USC