LECTURE 9:
Machine learning for paleoclimate

ML-4430: Machine learning approaches in climate science

22 June 2021

Climate Field Reconstructions

1

- Guillot, Rajaratnam & Emile-Geay, 2015

Paleoclimate Networks

2

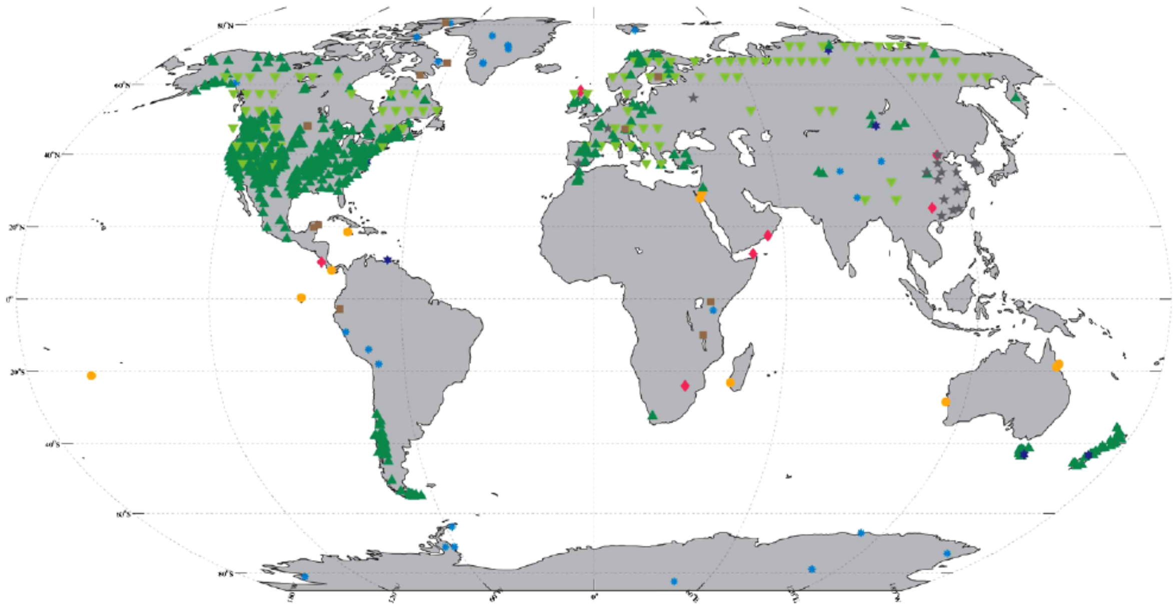
- Rehfeld, Molkenthin & Kurths, 2014

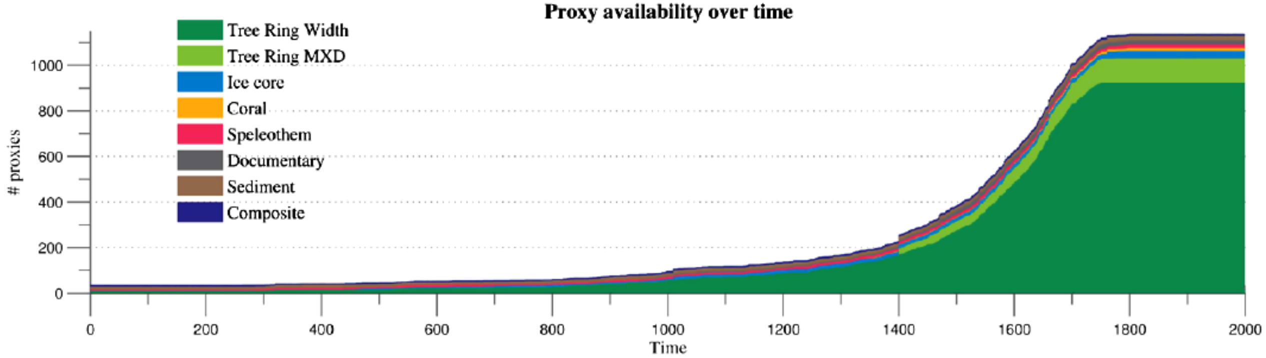
The Annals of Applied Statistics
2015, Vol. 9, No. 1, 324–352
DOI: 10.1214/14-AOAS794

STATISTICAL PALEOCLIMATE RECONSTRUCTIONS VIA MARKOV RANDOM FIELDS

BY DOMINIQUE GUILLOT[*,1,2],
BALA RAJARATNAM[*,2] AND JULIEN EMILE-GEAY[†,3]

Stanford University[*] and University of Southern California[†]





1. Guillot, Rajaratnam & Emile-Geay, 2015 → The challenge

Bedartha Goswami

Temperature and proxy as multivariate normal random variates

$$(X_1, X_2, \ldots, X_p) \sim N_p(\mu, \Sigma)$$

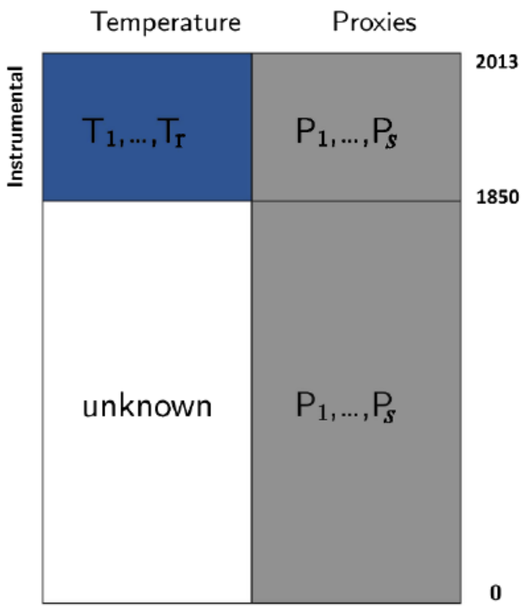
$$\mu = (\mu_1, \mu_2, \ldots, \mu_p)$$

$$\Sigma = (\sigma_{ij})_{pxp}$$

$$n = n_a + n_m$$

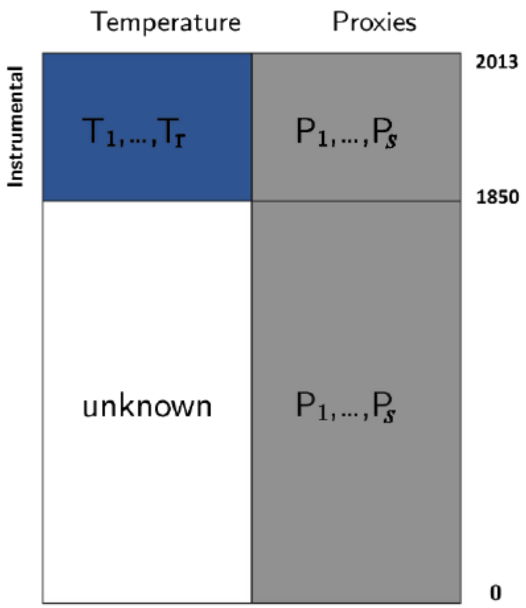
$$p \approx 3000, n \approx 2000, n_a = 150$$

where p → number of locations, and
n → number of years
n_a → number of years with available data
n_m → number of years with missing data



For large p, small n, sample covariance matrix is a poor estimator of Σ

$$p \approx 3000, n \approx 2000, n_a = 150$$



Bedartha Goswami

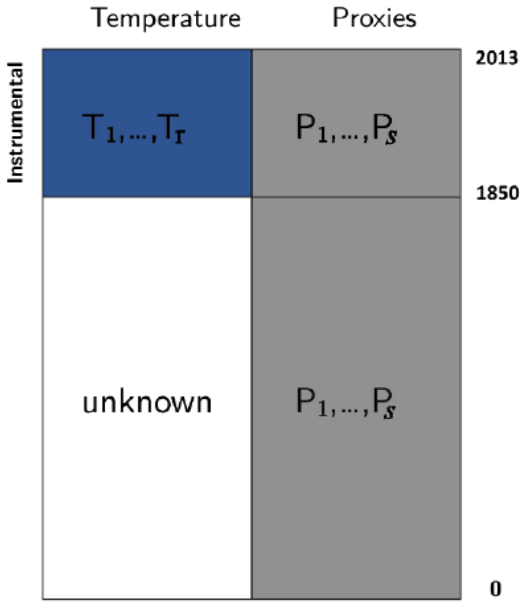
Expectation step: linear regression

$$(x_m - \mu_m^{(l)})^\top = B^{(l)}(x_a - \mu_a^{(l)})^\top, \tag{1.1}$$

where

$$B^{(l)} = \Sigma_{ma}^{(l)}(\Sigma_{aa}^{(l)})^{-1}, \qquad \Sigma^{(l)} = \begin{pmatrix} \Sigma_{aa}^{(l)} & \Sigma_{am}^{(l)} \\ \Sigma_{ma}^{(l)} & \Sigma_{mm}^{(l)} \end{pmatrix} \quad \text{and}$$

$$\mu^{(l)} = (\mu_a^{(l)}, \mu_m^{(l)}), \tag{1.2}$$



Here, *l* is the index for the current iteration number.

Initial guesses:

- $\mu^{(0)}$: sample mean, replacing all misssing values :
- $\Sigma^{(0)}$: sample covariance

1. Guillot, Rajaratnam & Emile-Geay, 2015 → The EM algorithm

Maximisation step: update of μ and Σ

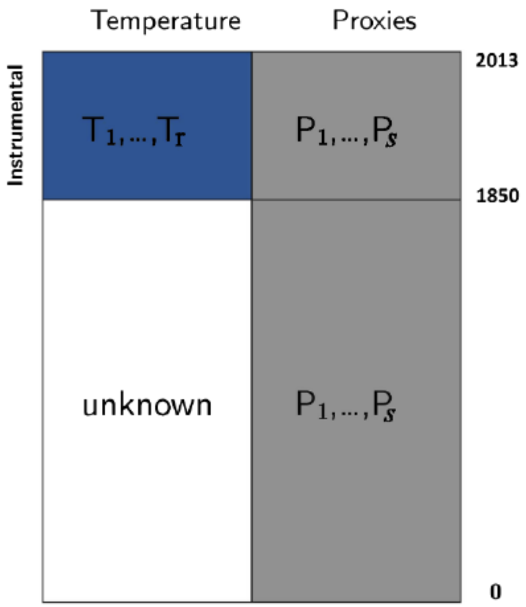
$$\mu_i^{(l+1)} = \frac{1}{n}\sum_{k=1}^{n} X_{ki}^{(l+1)},$$

(1.3)

$$\Sigma_{ij}^{(l+1)} = \frac{1}{n}\sum_{k=1}^{n}[(X_{ki}^{(l+1)} - \mu_i^{(l+1)})(X_{kj}^{(l+1)} - \mu_j^{(l+1)})] + C_{ij}^{(l+1)},$$

where $C_{ij}^{(l+1)}$ is the covariance of the residuals. Using the same block decomposition as in (1.2), we have

$$C^{(l+1)} = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{mm}^{(l)} - \Sigma_{ma}^{(l)}(\Sigma_{aa}^{(l)})^{-1}\Sigma_{am}^{(l)} \end{pmatrix}. \quad (1.4)$$



But we do not have a good estimate for Σ

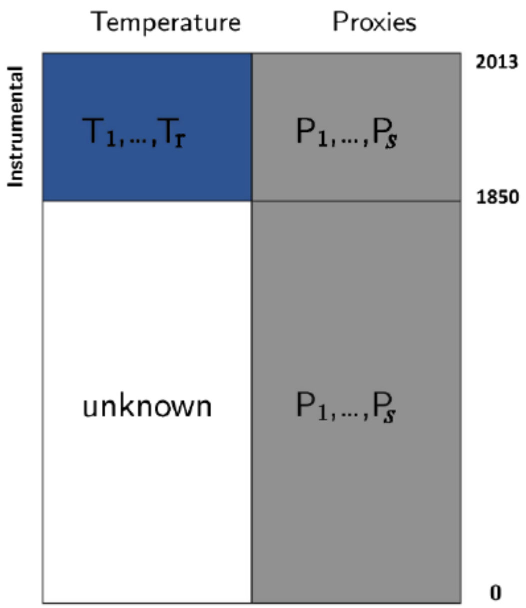
- Use regularised regression (RegEM & RegEM-TTLS)
- Here, we use Gaussian Markov Random Fields to estimate Σ
- First, consider the precision matrix,

  $$\Omega = (\omega_{ij}) = \Sigma-1$$

- Next, note that the partial correlation coefficient ρ_{ij} between *i* and *j* given all the other observables is given by

  $$\rho_{ij\,|\,\mathrm{rest}} = \quad -\,\omega_{ij} \,/\, (\omega_{ii}\,\omega_{jj})^{-1/2}$$

  i.e. X_i and X_j are independent given the rest of the data if the corresponding entry in Ω is zero,
  and vice versa!



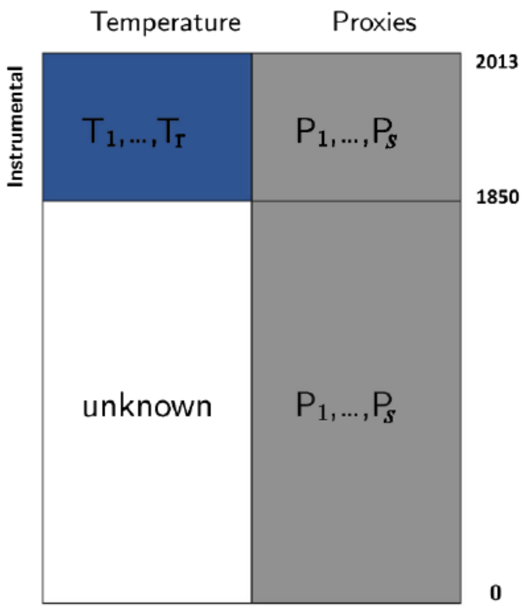
The graph-based estimator for Σ

$$\hat{\Sigma}_G = \underset{\substack{\Sigma=\Omega^{-1}>0 \\ \omega_{ij}=0,(i,j)\notin E}}{\operatorname{argmax}} \log\det\Omega - \operatorname{tr}(S\Omega), \tag{1.6}$$

where S is the sample covariance matrix of $x_1, \ldots, x_n$, given by

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^\top, \tag{1.7}$$

But we need to estimate Ω first



1. Guillot, Rajaratnam & Emile-Geay, 2015 → The GraphEM algorithm

Estimating Ω

$$\max_{\Omega>0} l(\Omega) - \rho\|\Omega\|_1, \tag{2.1}$$

where $\Omega = \Sigma^{-1}$ denotes the precision matrix of the data, $l(\Omega)$ is the normal log-likelihood of Ω, $\rho > 0$ is a regularization parameter, and $\|\Omega\|_1$ is the 1-norm of Ω:

$$\|\Omega\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} |\omega_{ij}|. \tag{2.2}$$

$$\max_{\Omega>0} l(\Omega) - \rho_{\mathrm{TT}}\|\Omega_{\mathrm{TT}}\|_1 - 2\rho_{\mathrm{TP}}\|\Omega_{\mathrm{TP}}\|_1 - \rho_{\mathrm{PP}}\|\Omega_{\mathrm{PP}}\|_1, \tag{2.4}$$

$$\Omega = \begin{pmatrix} \Omega_{\mathrm{TT}} & \Omega_{\mathrm{TP}} \\ \Omega_{\mathrm{PT}} & \Omega_{\mathrm{PP}} \end{pmatrix},$$

1. Guillot, Rajaratnam & Emile-Geay, 2015 → The GraphEM algorithm

Evaluatin GraphEM

- Pseudoproxy measurements
  - NCAR CSM 1.4 model experiment
  - Last millennium (850–1980 AD)
  - $5^\circ \times 5^\circ$ grid

$$P(l,t) = T(l,t) + \frac{1}{\mathrm{SNR}} \cdot \xi(l,t),$$

- Most recent 150 years used for calibration
- The 981 years before that are reconstructed using GraphEM
- Benchmark comparison: RegEM-TTLS

- Performance metrics
  - Mean Squared Error (MSE)
  - Reduction of Error (RE)
  - Coefficient of Efficiency (CE)

$$\mathrm{RE}(l) = 1 - \frac{\mathrm{MSE}(\hat{T})(l)}{\mathrm{MSE}(\overline{T}_c)(l)}.$$

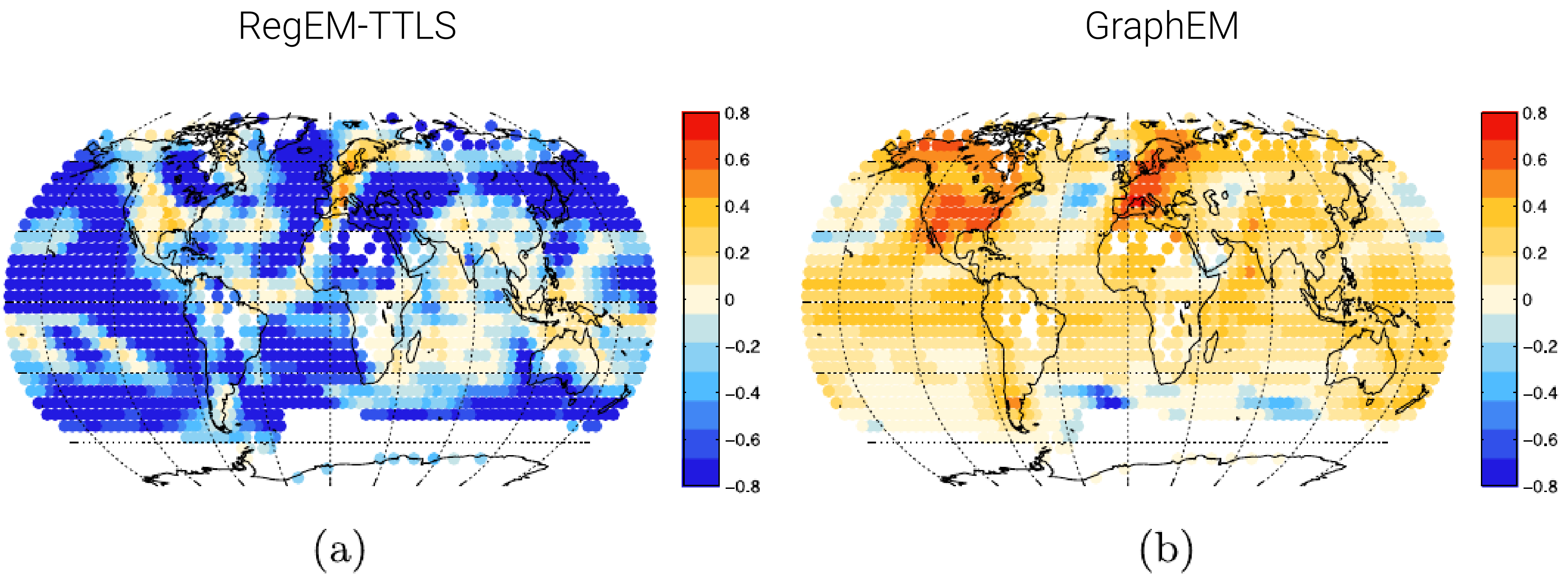
$$\mathrm{CE}(l) = 1 - \frac{\mathrm{MSE}(\hat{T})(l)}{\mathrm{MSE}(\overline{T}_v)(l)}.$$

1. Guillot, Rajaratnam & Emile-Geay, 2015 → Testing the GraphEM algorithm

Method	MSE	RE	CE	Bias
ℓ_1 method				
GraphEM (0.3% target sparsity)	0.44 (0.01)	0.33 (0.01)	0.11 (0.02)	0.09 (0.01)
GraphEM (0.5% target sparsity)	0.42 (0.01)	0.36 (0.01)	0.15 (0.01)	0.08 (0.01)
GraphEM (0.7% target sparsity)	0.41 (0.01)	0.36 (0.01)	0.16 (0.01)	0.08 (0.01)
GraphEM (0.9% target sparsity)	0.41 (0.01)	0.36 (0.01)	0.15 (0.01)	0.08 (0.01)
Neigh				
GraphEM (600 km radius)	0.42 (0.01)	0.35 (0.01)	0.14 (0.01)	0.06 (0.01)
GraphEM (800 km radius)	0.39 (0.01)	0.39 (0.01)	0.19 (0.01)	0.06 (0.01)
GraphEM (1000 km radius)	0.40 (0.01)	0.38 (0.01)	0.18 (0.01)	0.06 (0.01)
GraphEM (1200 km radius)	0.41 (0.01)	0.36 (0.01)	0.16 (0.01)	0.06 (0.01)
RegEM-TTLS				
RegEM-TTLS	0.84 (0.10)	−0.24 (0.14)	−0.61 (0.19)	0.01 (0.02)

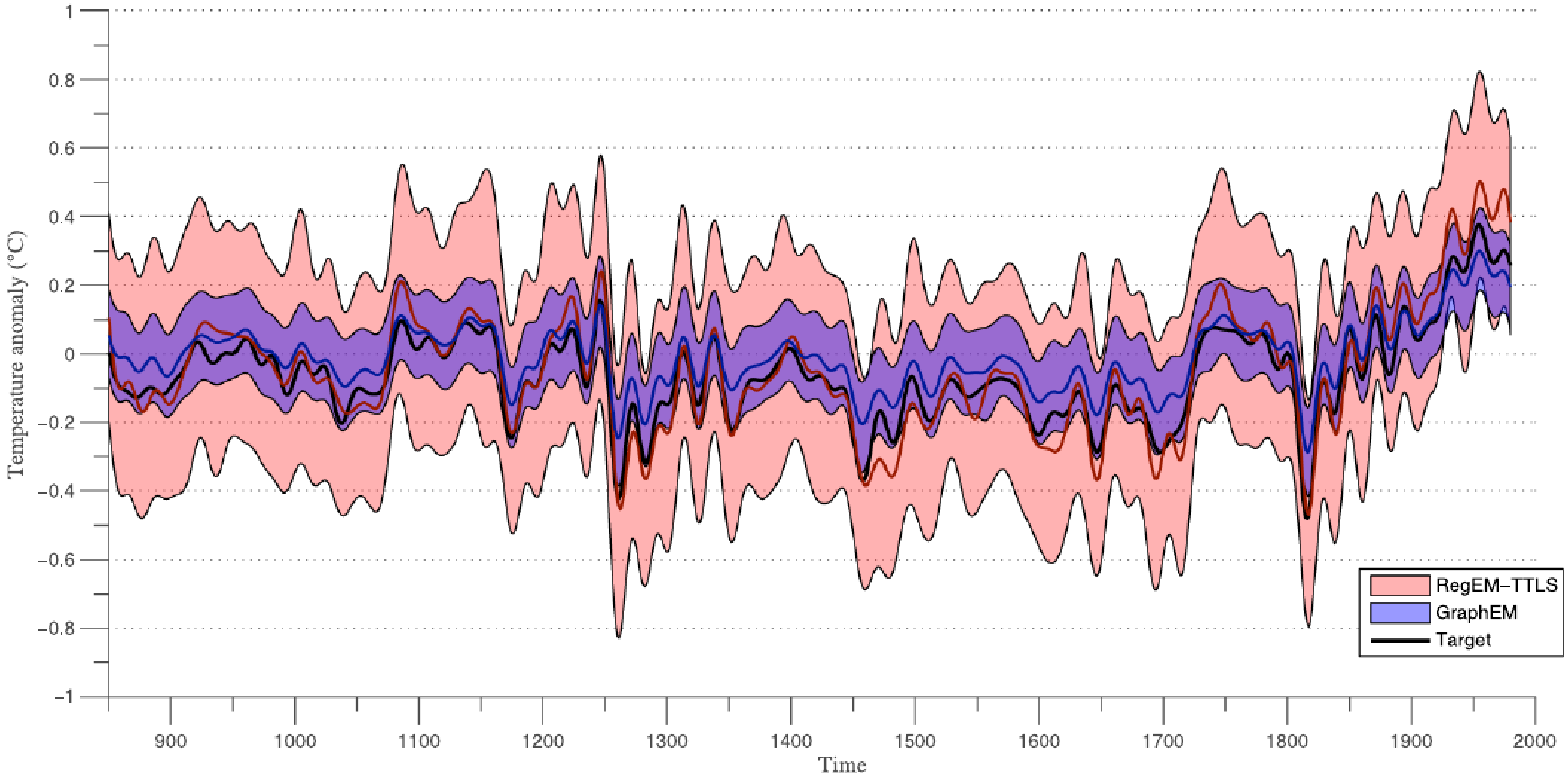
1. Guillot, Rajaratnam & Emile-Geay, 2015 → Results

CE



1. Guillot, Rajaratnam & Emile-Geay, 2015 → Results

Global Average Temperature



1. Guillot, Rajaratnam & Emile-Geay, 2015 → Results

Climate Field Reconstructions

1. Guillot, Rajaratnam & Emile-Geay, 2015

Paleoclimate Networks

2. Rehfeld, Molkenthin & Kurths, 2014

Nonlin. Processes Geophys., 21, 691–703, 2014
www.nonlin-processes-geophys.net/21/691/2014/
doi:10.5194/npg-21-691-2014

Nonlinear Processes in Geophysics
Open Access

Testing the detectability of spatio–temporal climate transitions from paleoclimate networks with the START model

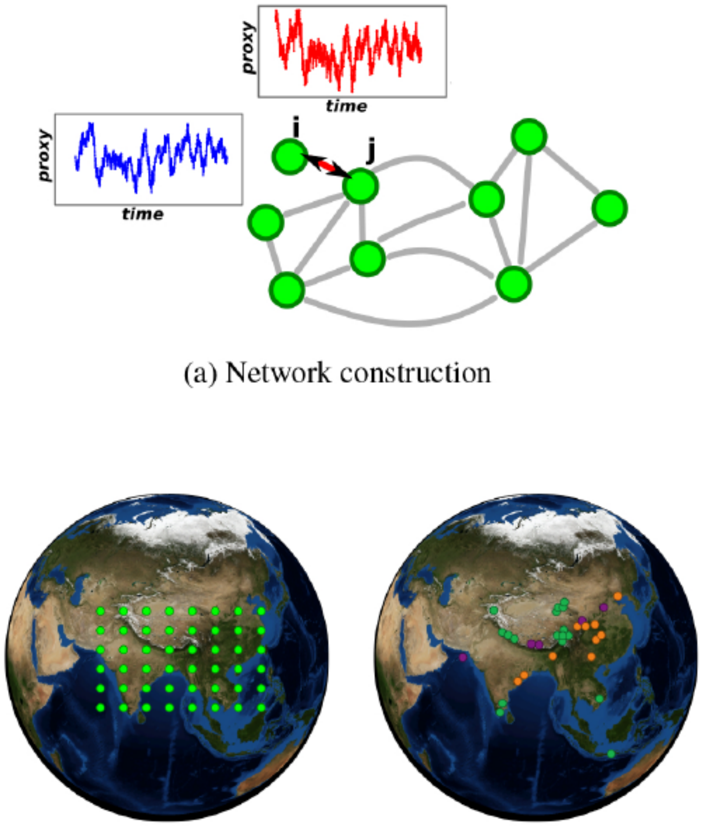
K. Rehfeld[1,2,3], **N. Molkenthin**[1,2], **and J. Kurths**[1,2]

[1]Potsdam Institute for Climate Impact Research, P.O. Box 601203, 14412 Potsdam, Germany
[2]Department of Physics, Humboldt-Universität zu Berlin, Newtonstr. 15, 12489 Berlin, Germany
[3]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Telegrafenberg A43, 14473 Potsdam, Germany

2. Rehfeld, Molkenthin & Kurths, 2014

Challenge ...

- Paleoclimate networks can be constructed from sparsely distributed proxy time series
- How well do they capture the dynamics of the system?
- If we had all the data from the region, and constructed a climate network, would the results be the same?



(a) Network construction

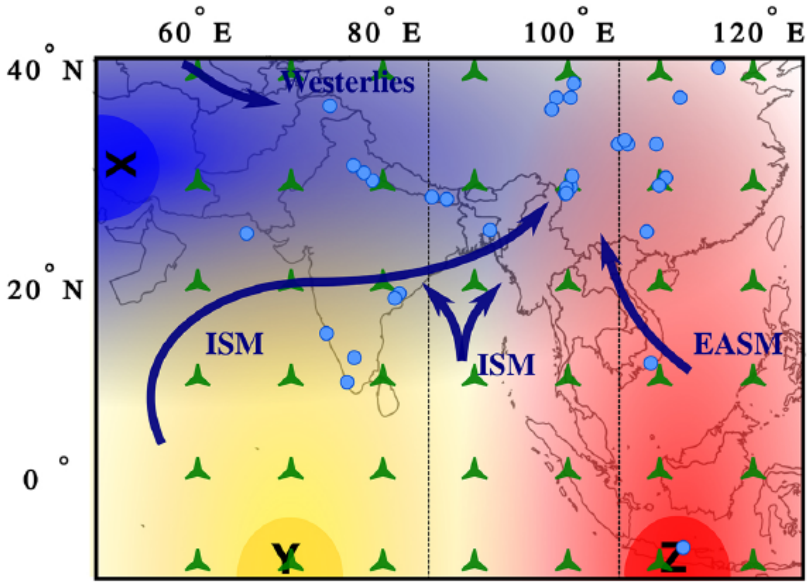
(b) Climate network

(c) Paleoclimate network

2. Rehfeld, Molkenthin & Kurths, 2014

START model

- Spatiotemporal autocorrelated time series (START) model
- From expert knowledge, consider 3 sources of wind (perturbations) in the Asian monsoon domain:
  - Westerlies (X)
  - ISM (Y)
  - EASM (Z)
- At each source, place a Gaussian wave front propagating through the domain using the advection-diffusion equation
- Innately, model internal dynamics at each location as a autoregressive process



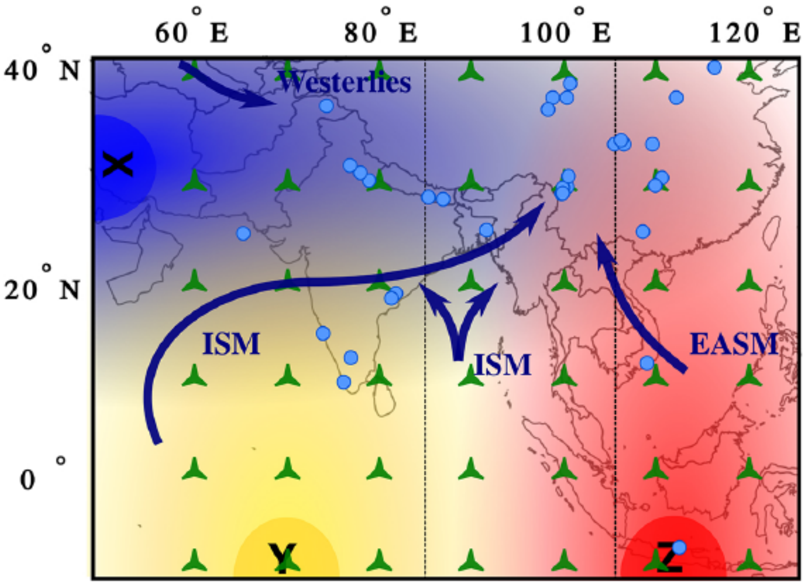
unidirectional front with a velocity at position $\boldsymbol{p}$ and time point t

$$v_X(t, \boldsymbol{p}, m_X, W) = m_X\, e^{-(p_x - p_{0,x})^2/2W}, \tag{14}$$

with a full width at a half maximum of $2\sqrt{W \log 2}$. The maximal amplitude of the velocity, m_X,

$$m_X(t, F) = m_X = B_X + \alpha F, \tag{15}$$

is found in the center of the Gaussian front, as in Fig. 3. Here, B_X is the baseline strength of the component's flow, and α is its amplitude, or susceptibility to the external forcing, repre-



2. Rehfeld, Molkenthin & Kurths, 2014 → START model

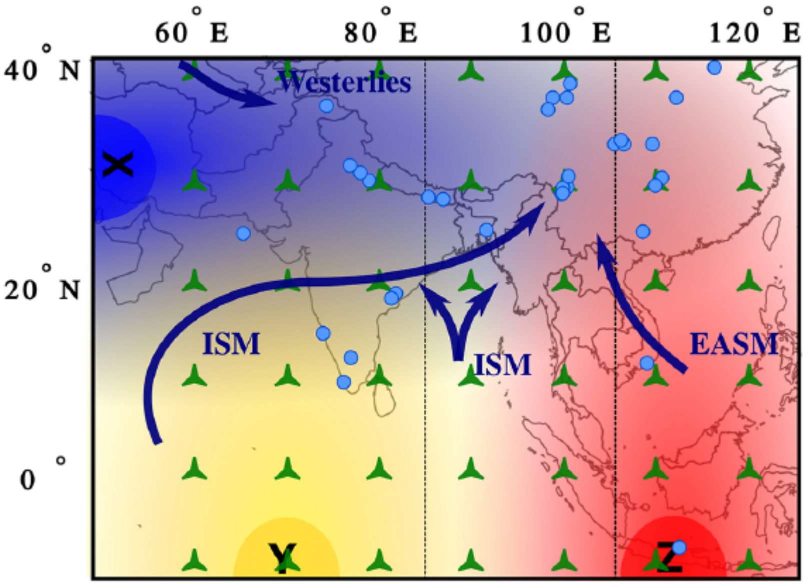
Signal at point i

ITCZ component

$$R_i = f_X(F,i)R_X + f_Y(F,i)R_Y + f_Z(F,i)R_Z + R_{\text{noise}} \quad (16)$$

Westerlies component

EASM component



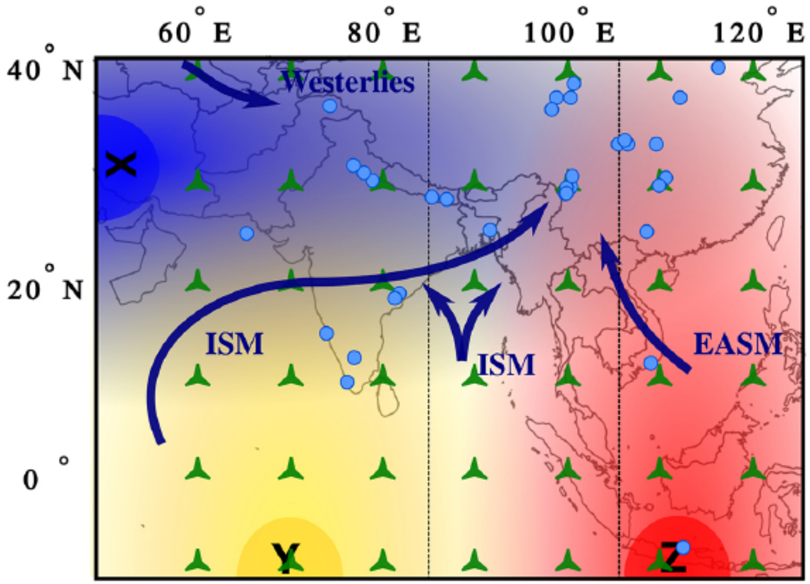
2. Rehfeld, Molkenthin & Kurths, 2014 → START model

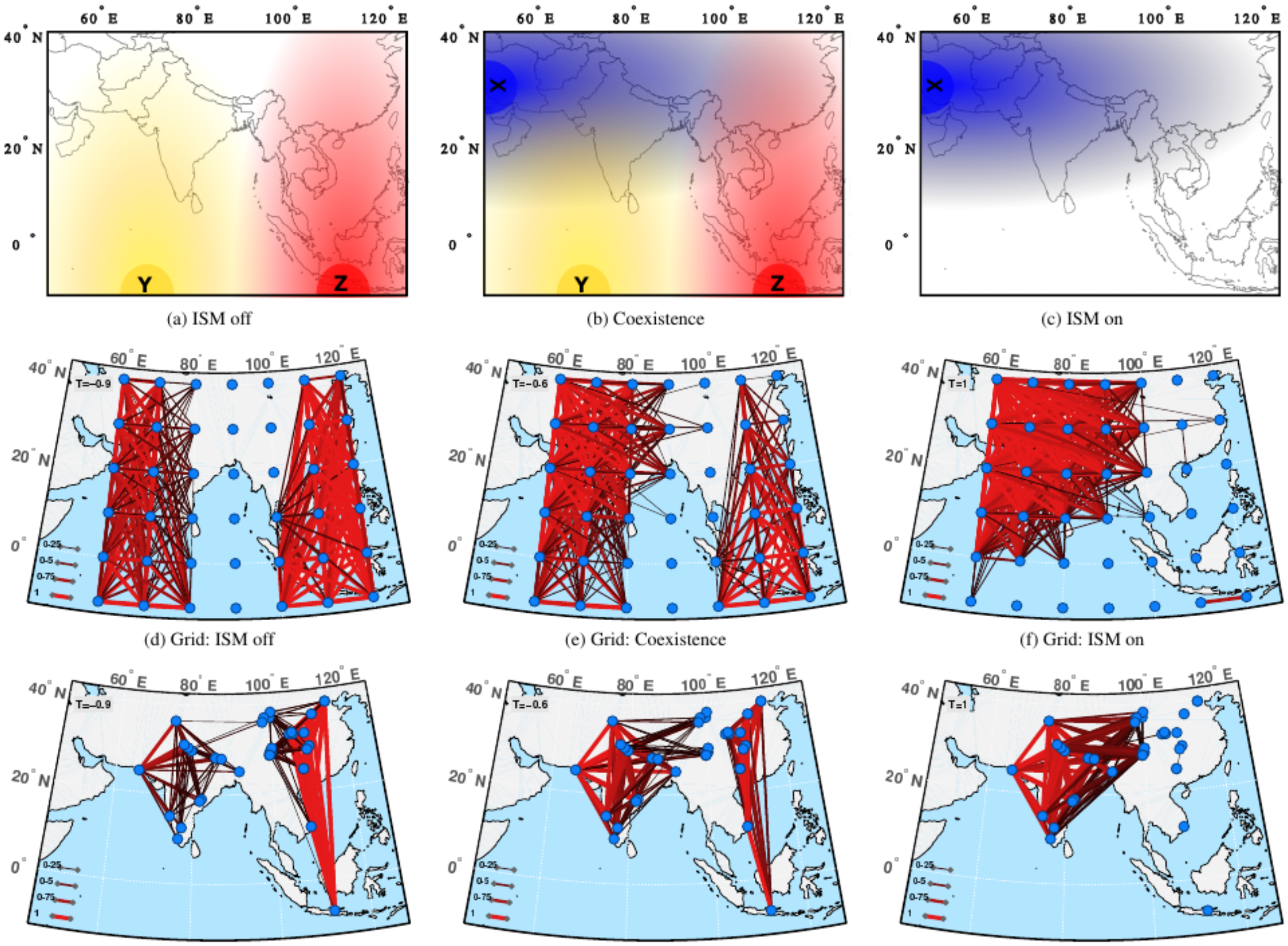
Experimental setup

- Model runs are sampled in two different ways:
  - In the form of a regular grid
  - Locations of paleoclimate records from previous studies

- Three different experiments:
  - ISM off: Westerlies source forcing is removed
    => F = -1
  - Coexistence: All three sources are kept
    => F = 0
  - ISM on: Sources Y and Z are removed
    => F = 1

- Climate networks are constructed:
  - From the gridded samples
  - From the paleoclimate record locations samples



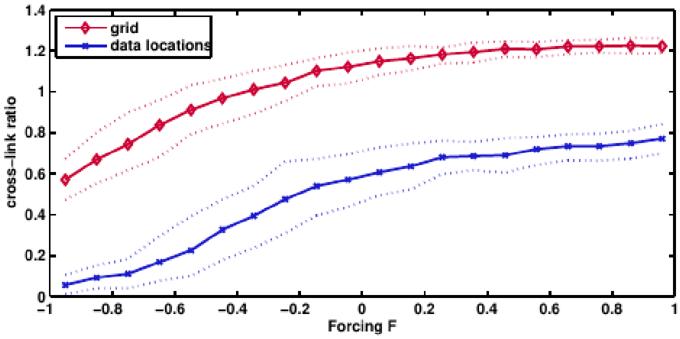


(a) ISM off (b) Coexistence (c) ISM on

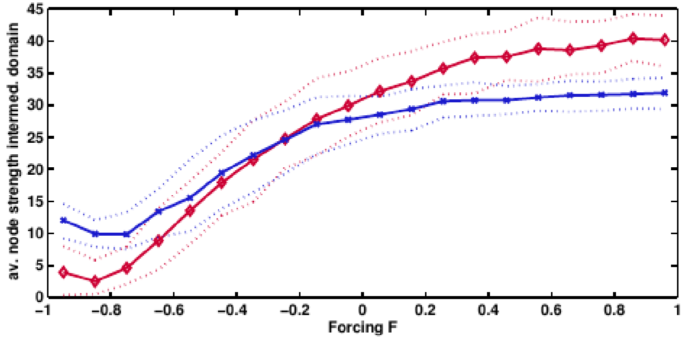
(d) Grid: ISM off (e) Grid: Coexistence (f) Grid: ISM on

(g) Data locations: ISM off (h) Data locations: Coexistence (i) Data locations: ISM on

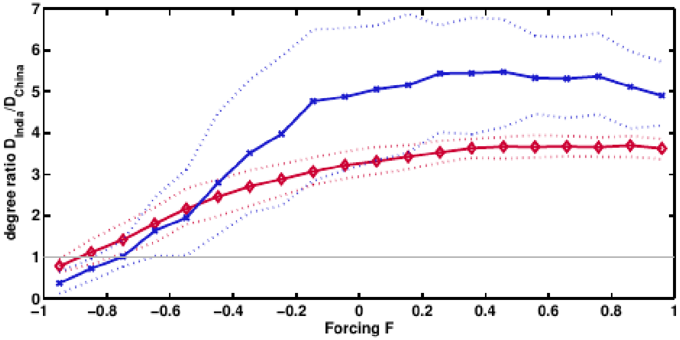
2. Rehfeld, Molkenthin & Kurths, 2014 → Results



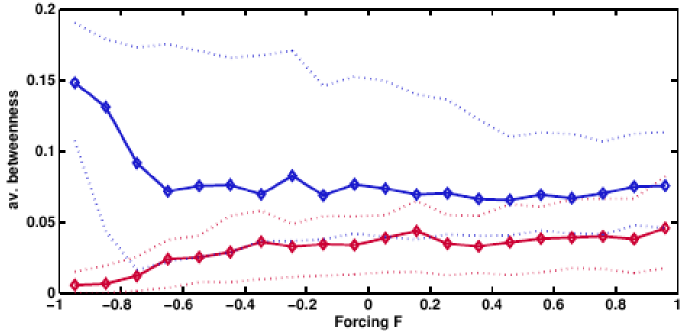
(a) Cross link ratio



(b) Average node strength



(c) Regional degree ratio



(d) Shortest path betweenness

2. Rehfeld, Molkenthin & Kurths, 2014 → Results

Climate Field Reconstructions

1

- Guillot, Rajaratnam & Emile-Geay, 2015

Paleoclimate Networks

2

- Rehfeld, Molkenthin & Kurths, 2014