Probabilistic Graphical Models for Climate Data Analysis

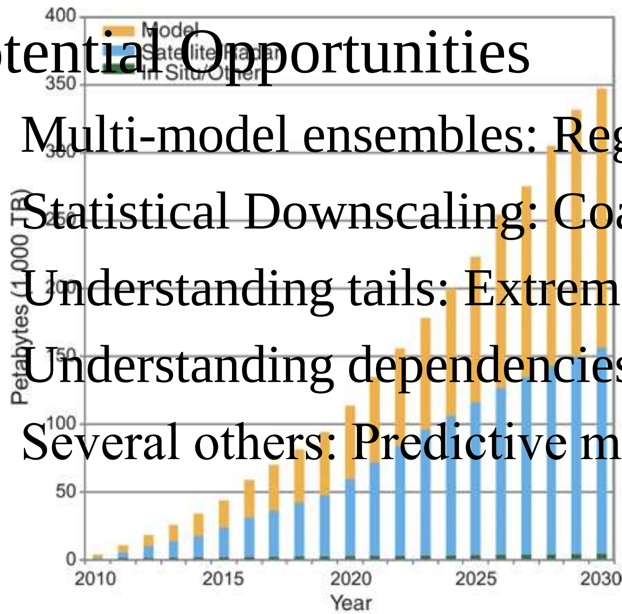
Arindam Banerjee
banerjee@cs.umn.edu

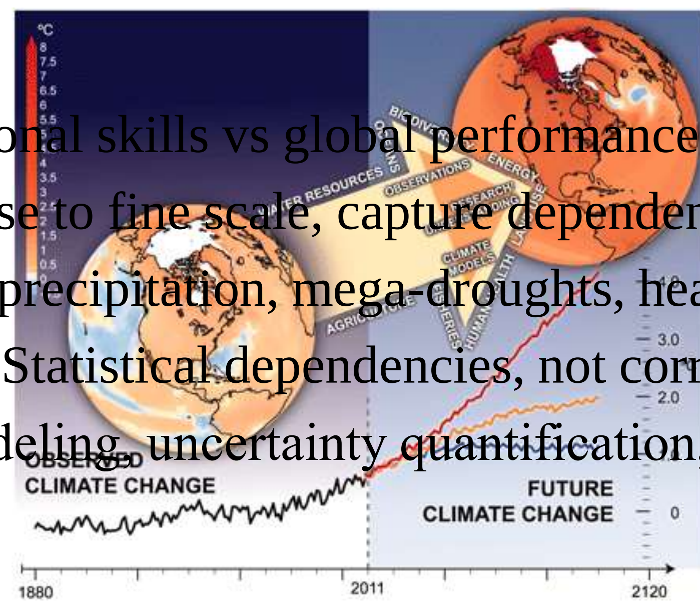
Dept of Computer Science & Engineering
University of Minnesota, Twin Cities

Aug 15, 2013

Climate Data Analysis

- Key Challenges
  - High-dimensional dependent data, small sample size
  - Spatial and temporal dependencies, temporal lags
  - Oscillations with frequency and phase variations
  - Important variables are unreliable, e.g., precipitation
  - Several others: Nonlinearity, heavy tails, …
- Potential Opportunities
  - Multi-model ensembles: Regional skills vs global performance
  - Statistical Downscaling: Coarse to fine scale, capture dependencies
  - Understanding tails: Extreme precipitation, mega-droughts, heat waves, etc.
  - Understanding dependencies: Statistical dependencies, not correlation
  - Several others: Predictive modeling, uncertainty quantification, …

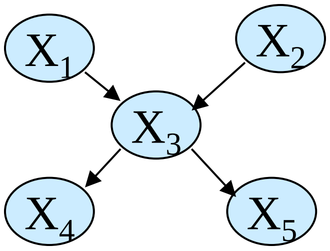




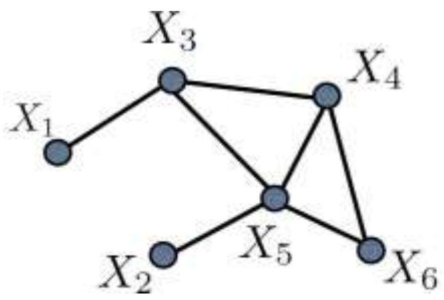
Source: Overpeck et al., *Science*, (2011)

Graphical Models

- Graphical models
  - Dependencies between (random) variables, avoid I.I.D. assumptions
  - Closer to reality, learning/inference is much more difficult
- Basic nomenclature
  - Node = Random Variable, Edge = Statistical Dependency
- Directed Graphs
  - A *directed* graph between random variables
  - Example: Bayesian networks, Hidden Markov Models
  - Joint distribution is a product of P(child|parents)

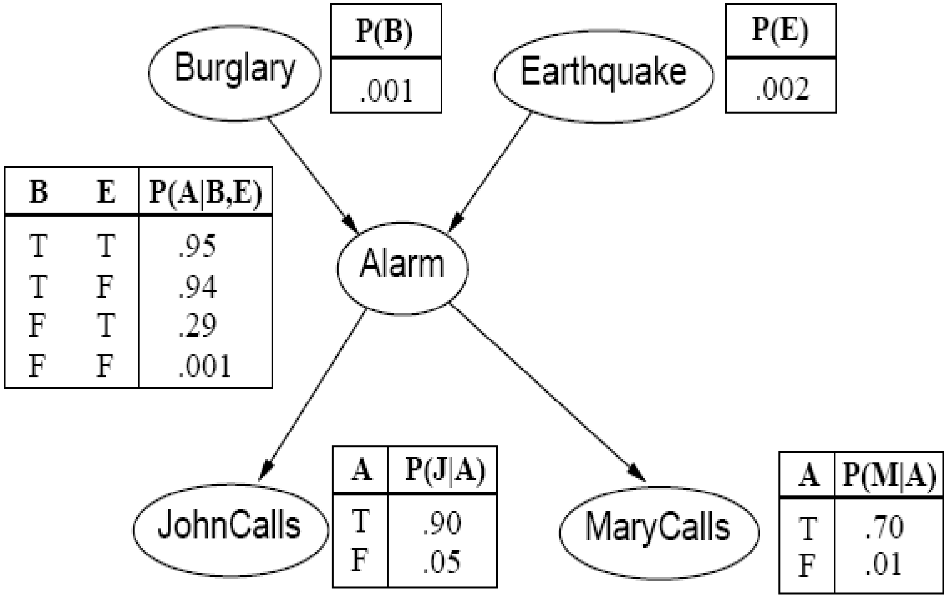


- Undirected Graphs
  - An *undirected* graph between random variables
  - Example: Markov/Conditional random fields
  - Joint distribution in terms of potential functions

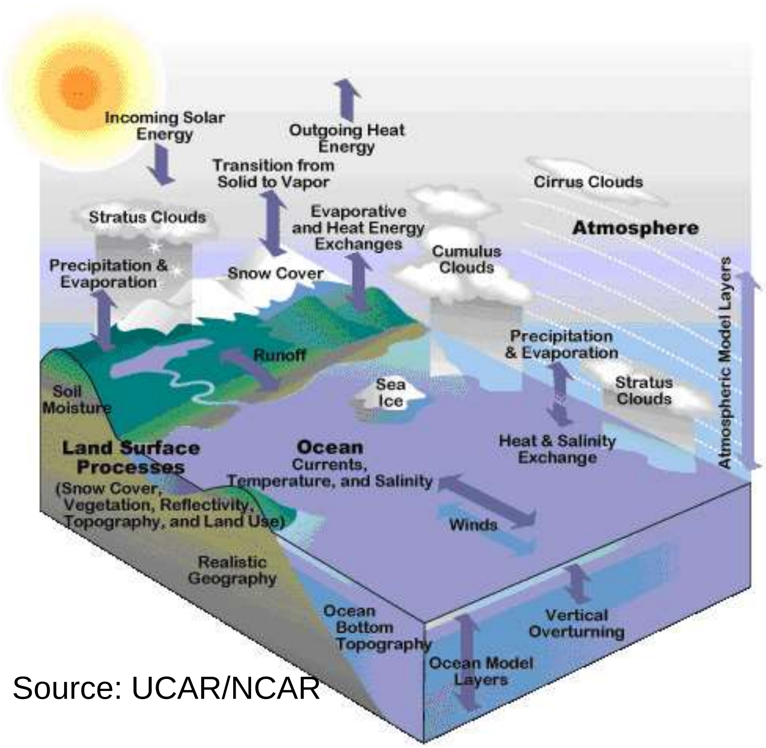


Graphical Models: Key Problems

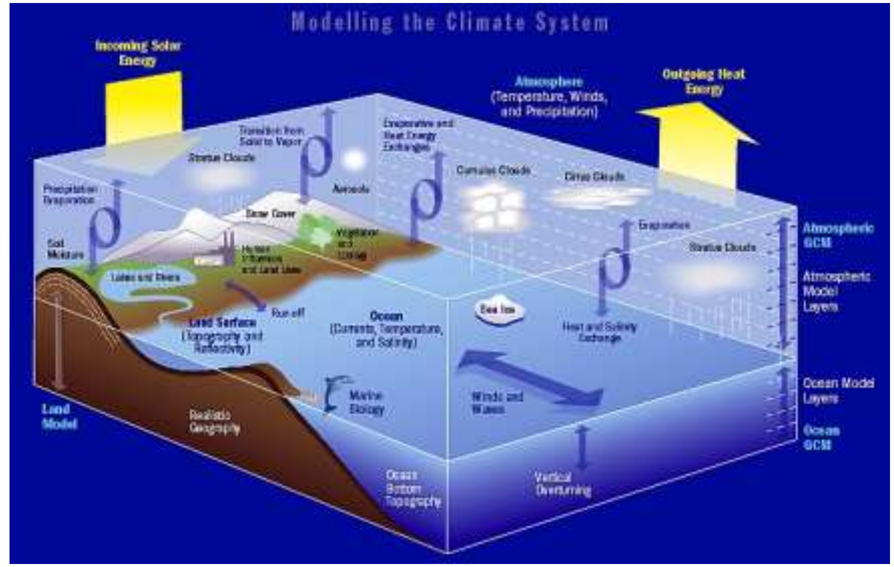
- Structure Learning
  - Given: Samples
  - Problem: Learn the Structure
- Parameter Estimation
  - Given: Samples and Structure
  - Problem: Estimate Parameters
- Inference
  - Given: Structure, Parameters, and some variables (part of a Sample)
  - Problem: Find other variables (part of a Sample)



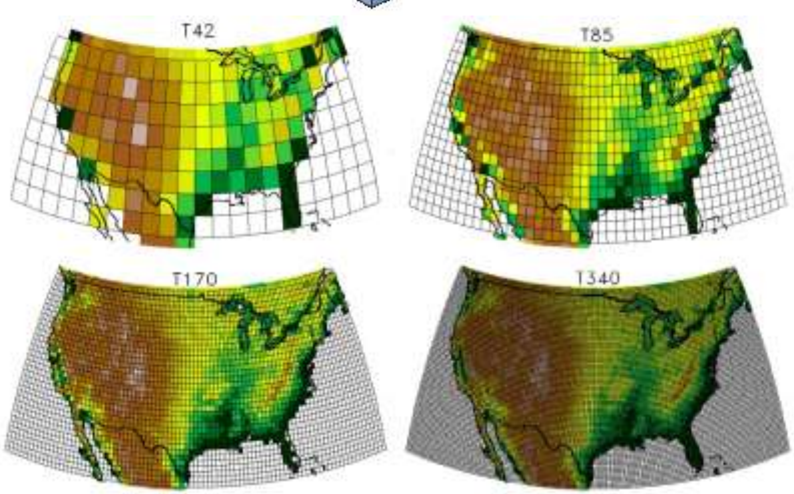
Global Climate Models (GCMs)

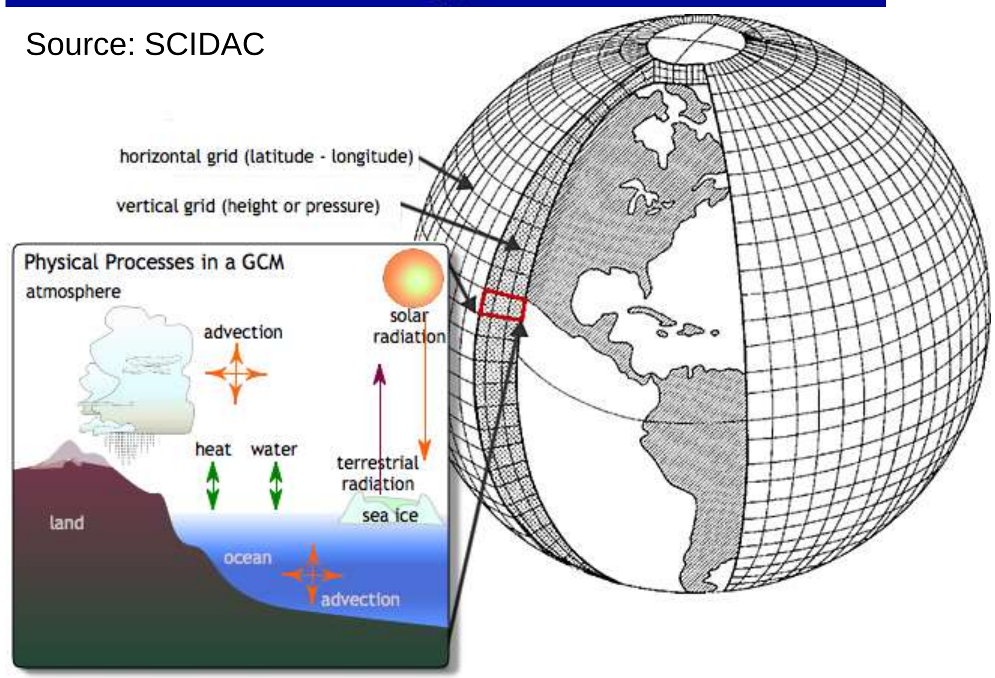


Source: UCAR/NCAR



Source: SCIDAC

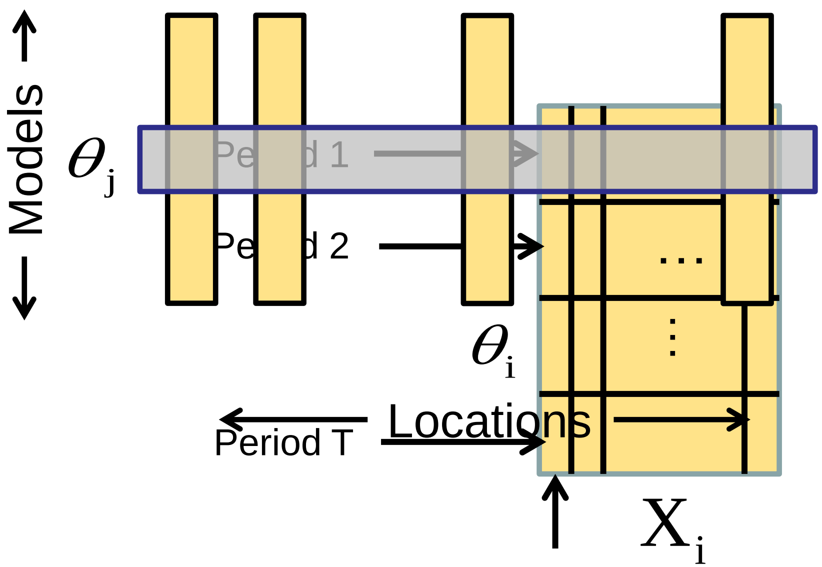




Combining GCM Outputs

- Several ways to combine the model outputs
  - Average: Equal weightage to all models (IPCC AR4 2007, Reifen and Toumi 2009)
  - Superensemble: Least Squares (Krishnamurti et al., 2002)
  - REA: Reliability based ensemble averaging (Giorgi et al., 2002)
  - Bayesian: Probabilistic estimates of climate variables (Tebaldi et al., 2005, Smith et al., 2011)
  - Online Learning: Tracking climate models (Monteleoni et al., 2011)
- Our work
  - Hypothesis: Certain models do well in certain climatic conditions
  - Goal: Climate model combination
    - Different weighs at different locations
    - Similar climatic conditions should get similar weights
  - Builds on superensemble and probabilistic approaches

Smooth Model Combination (SMC)

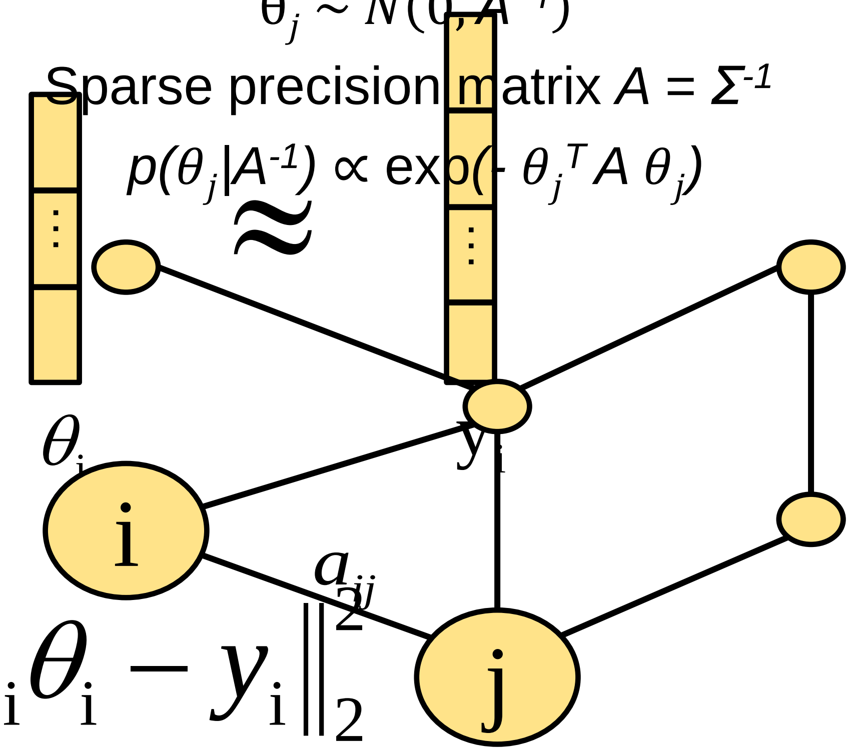


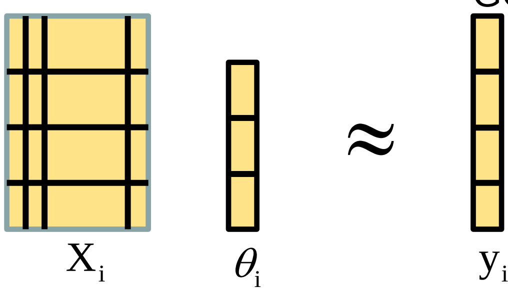
Prior to ensure smoothness

$\theta_j \sim N(0, A^{-1})$

Sparse precision matrix $A = \Sigma^{-1}$

$p(\theta_j|A^{-1}) \propto \exp(- \theta_j^T A \theta_j)$





$$\min_{\theta_i} \frac{1}{2} \left\| X_i \theta_i - y_i \right\|_2^2$$

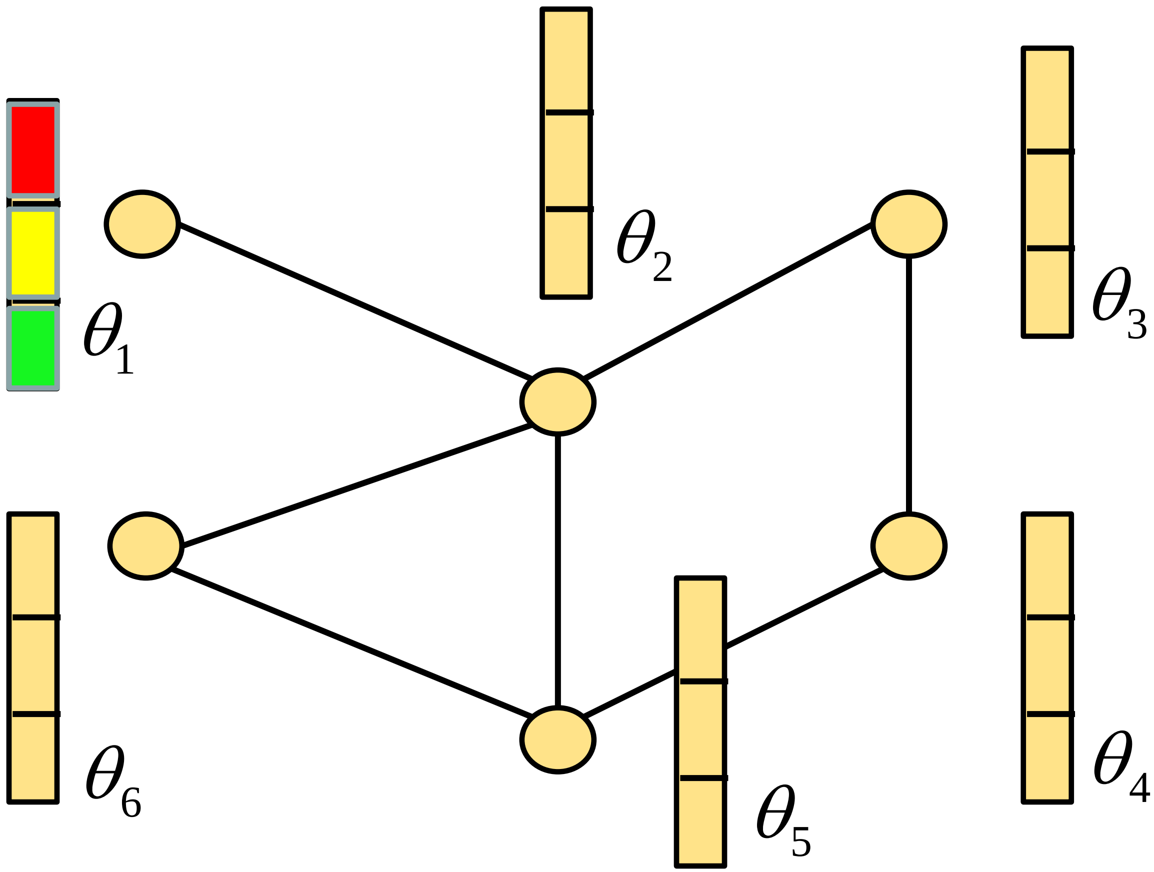
$$\min_{\theta_i} \frac{1}{2} \left\| X_i \theta_i - y_i \right\|_2^2$$

$$\min_{\Theta} \frac{1}{2} \sum_{i \in V} \left\| X_i \theta_i - y_i \right\|_2^2 + \frac{\lambda}{2} Tr(\Theta A \Theta^T)$$

SMC: Error Term

$$\min_{\Theta} \frac{1}{2} \sum_{i \in V} \left\| X_i \theta_i - y_i \right\|_2^2 + \frac{\lambda}{2} Tr(\Theta L \Theta^T)$$

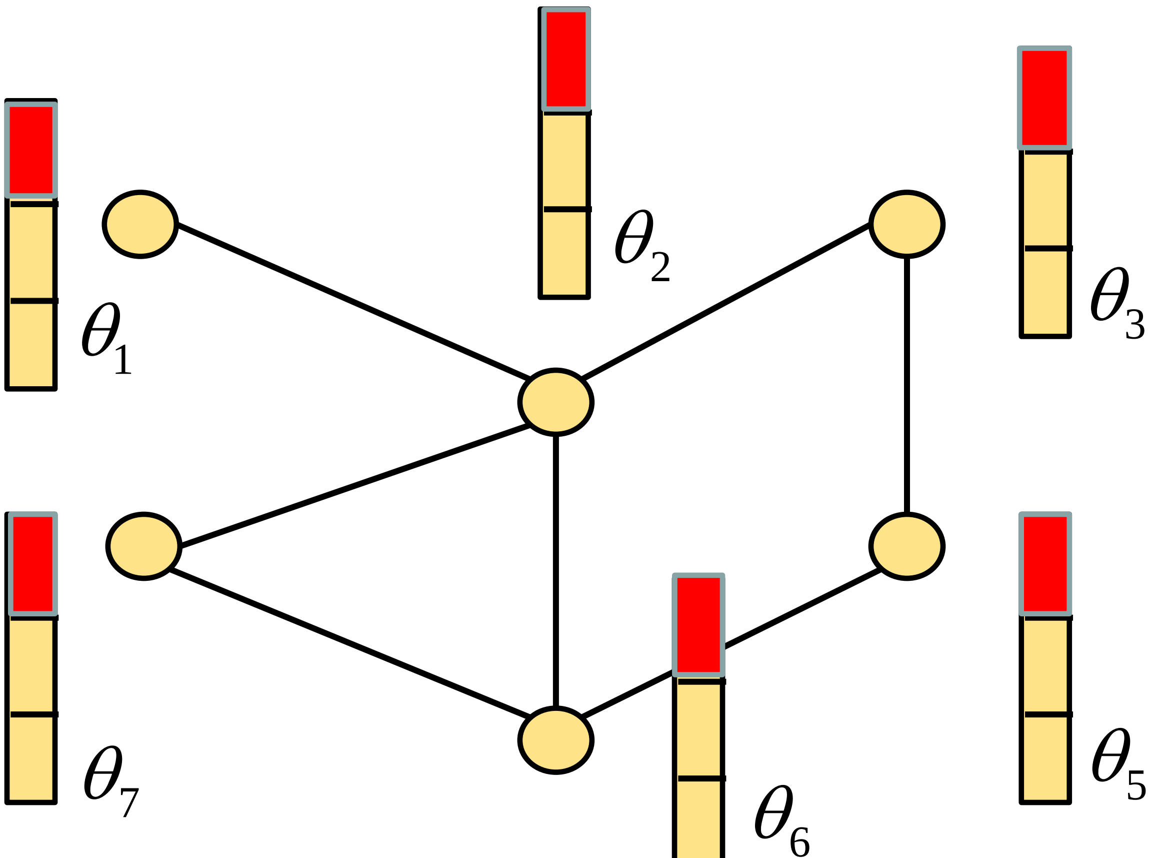
Error term updates locally



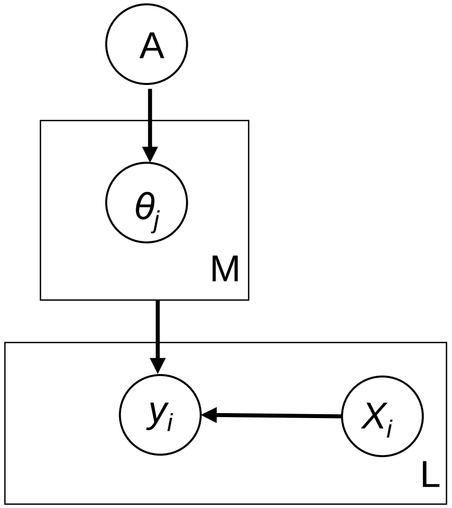
SMC: Smoothness Term

$$\min_{\Theta} \frac{1}{2} \sum_{i \in V} \left\| X_i \theta_i - y_i \right\|_2^2 + \frac{\lambda}{2} Tr(\Theta L \Theta^T)$$

Smoothness term updates based on neighborhoods



SMC: Graphical Model Perspective



- Generative Model

  Prior on rows: $\theta_j \sim N(0, A^{-1})$

  Conditional: $y_i \sim N(X_i \, \theta_i \, , \sigma^2)$

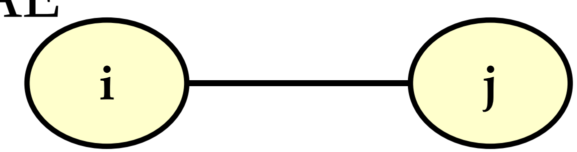
- Precision matrix specification
  - Gaussian Markov random field (GMRF)
  - Precision $A = L = D - W$, the discrete graph Laplacian
  - Intrinsic Conditionally Autoregressive Model (ICAR)
  - Spatial statistics literature (Diggle et al., 1998, Besag et al., 1995, Banerjee et al., 2004, Rue et al., 2005)

- Estimation of precision matrix
  - Estimate which locations are ‘similar’
  - Estimated precision is full rank but sparse

Data Set and Methodology

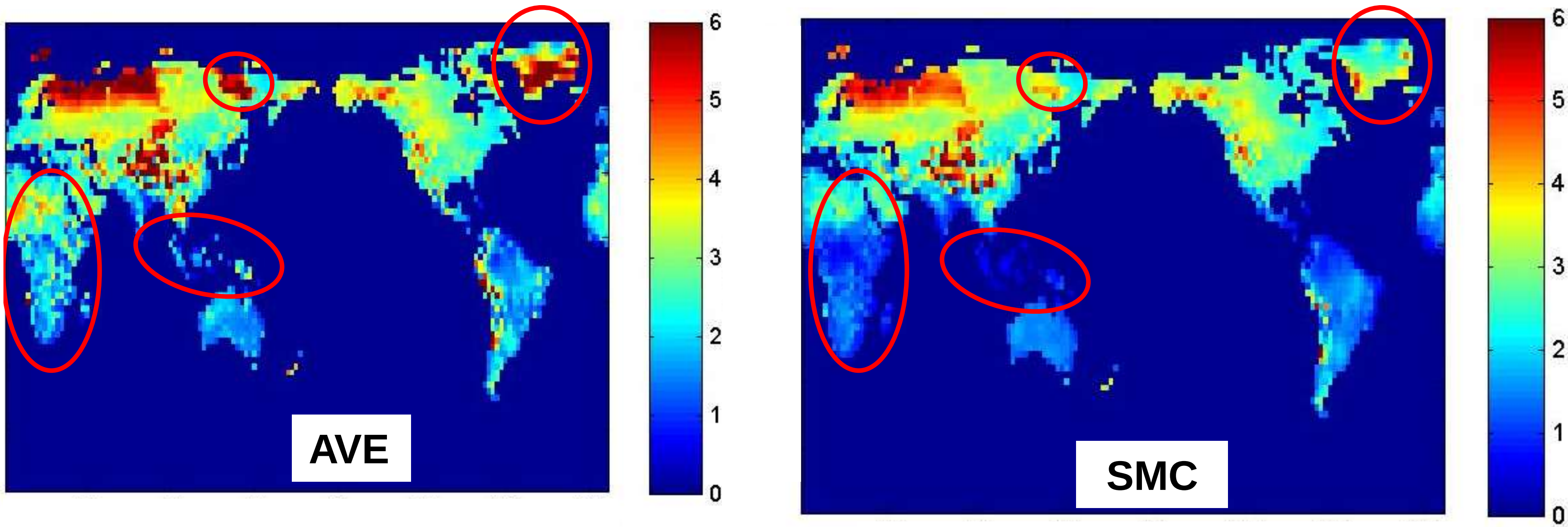
GCM	Origin
BCCR_BCM2	Bjerknes Centre for Climate Research, Norway
CCMA_CGCM3	Canadian Centre for Climate Modelling and Analysis, Canada
MICRO3-2-Hires	Center for Climate System Research, Univ. of Tokyo, Japan
CNRM_CM3	Center for National Weather Research, France
GFDL_CM2_1	Geophysical Fluid Dynamics Laboratory, USA
GISS_E_H	Goddard Institute for Space Studies, USA
INGV_ECHAM4	European Center for Medium-Range Weather Forecasts, UK
IPSL	Institut Pierre Simon Laplace, France

- GCM output: Monthly average surface temperature
- Target variable: Temperature from Climatic Research Unit (CRU)
- Error/accuracy measures: RMSE and MAE
- Smoothness measures:
  - Kendall τ
  - Spearman ρ



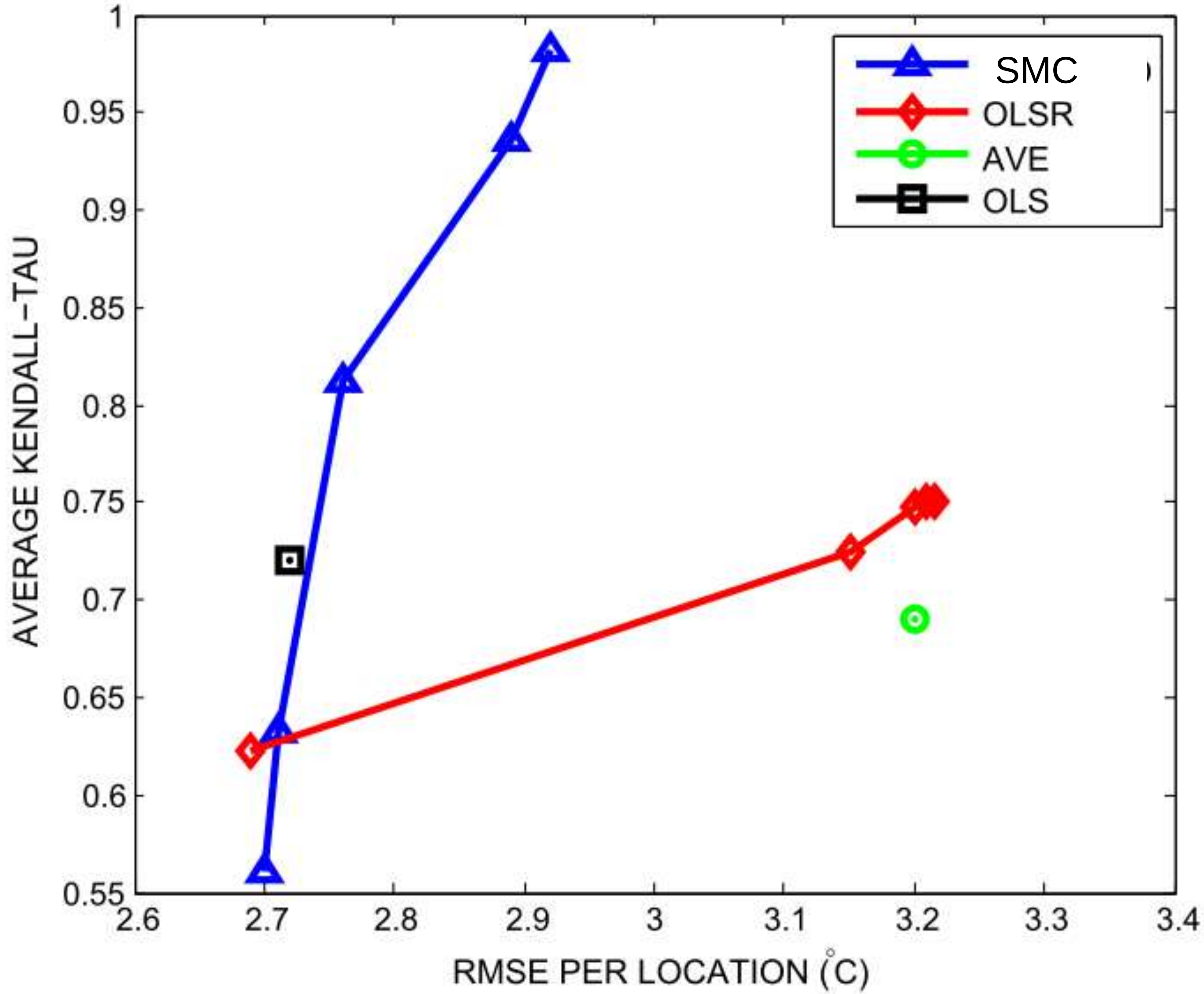
Sorted order of the regression coefficients ↓	i	j
	GCM1	GCM2
	GCM2	GCM1
	GCM3	GCM3
	GCM4	GCM4

Spatial Error Profile: AVE vs SMC



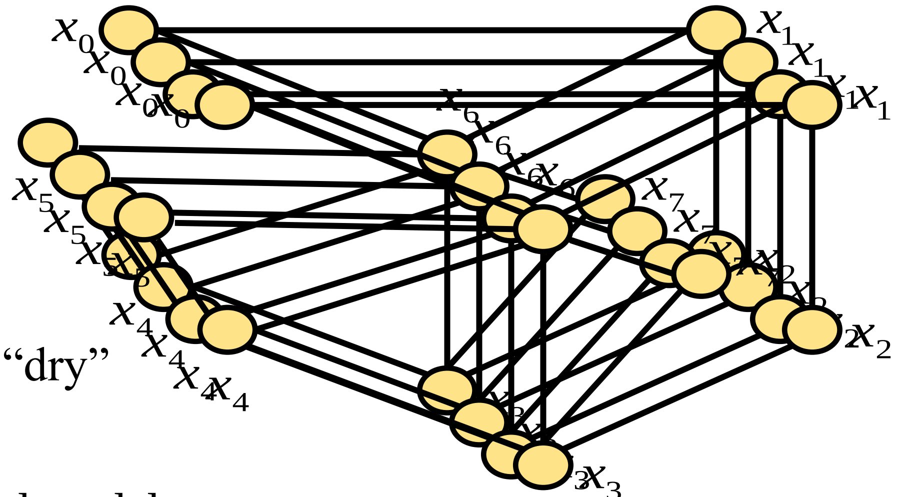
- SMC has lower compared to AVE: lower by ~ 0.5^0C
- Errors (visibly) reduced in many regions
  - Africa, Greenland, Southeast Asia, Siberia
- High errors in some regions
  - Northern Europe/Russia, China/Tibet, West South America, North America

Error vs Smoothness

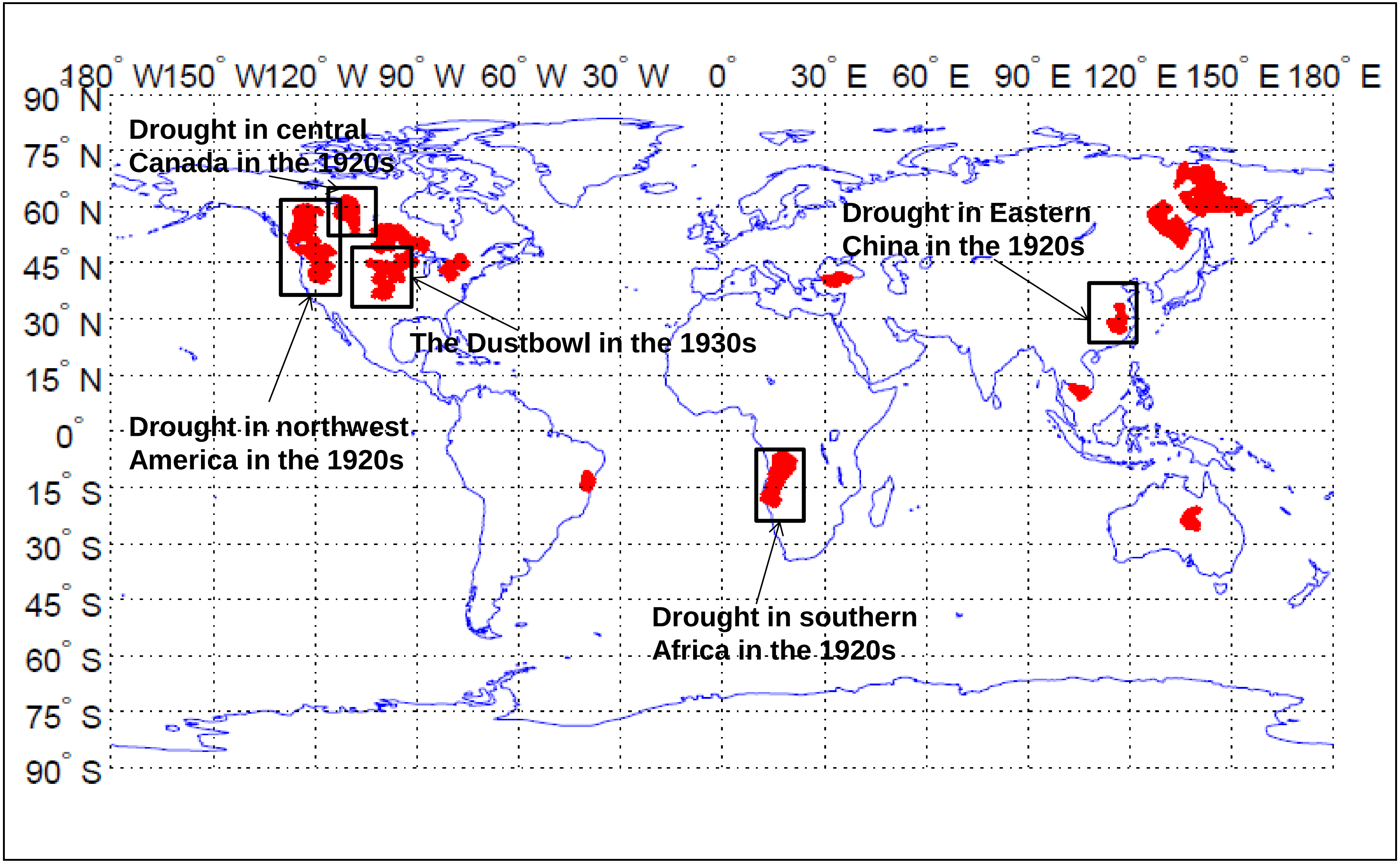


Mega-Droughts

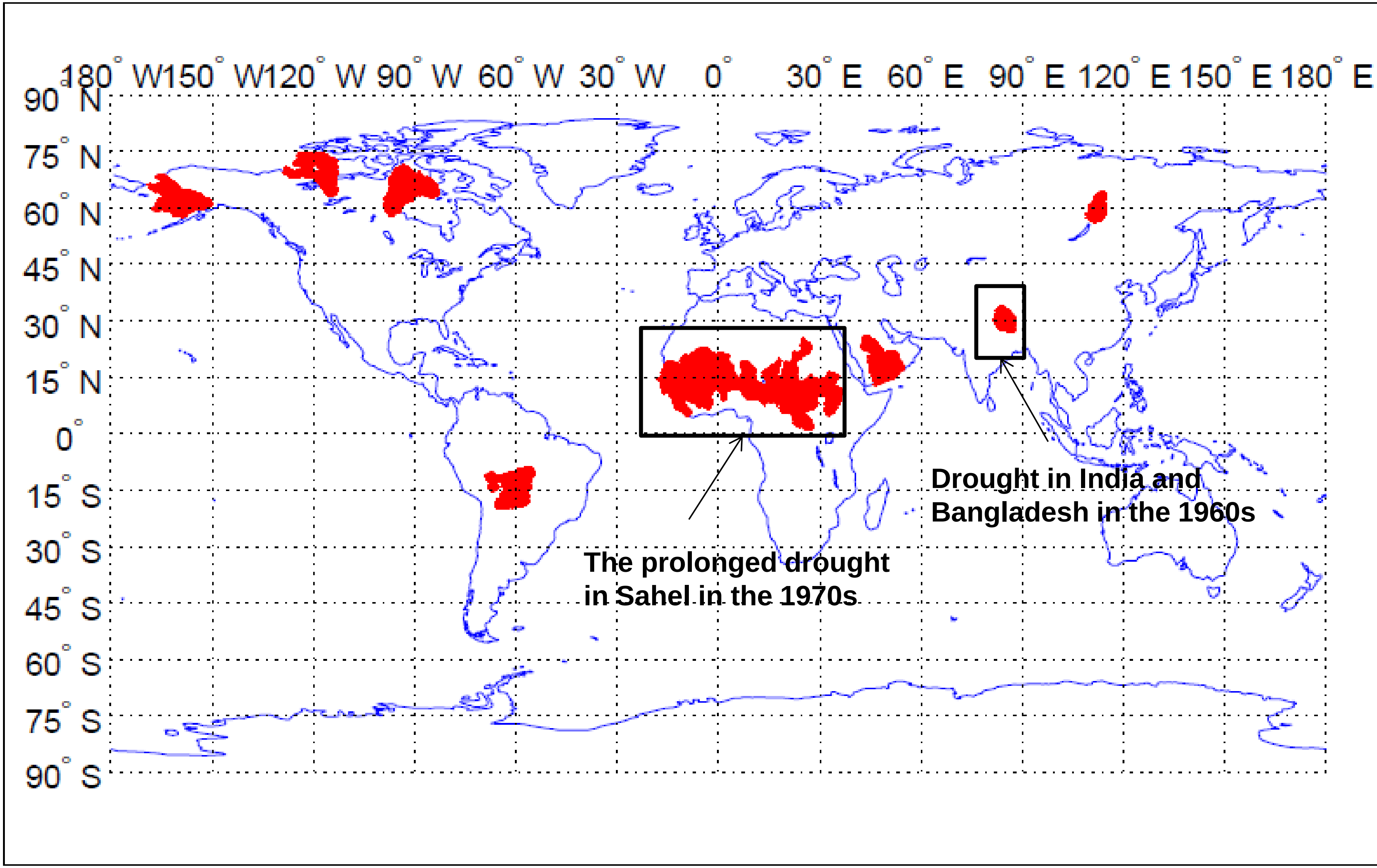
- Mega-Droughts
  - Persistent over space and time
  - Catastrophic consequences
- Examples
  - Late 1906s Sahel drought
  - 1930s North American Dust Bowl
- Discrete Markov Random Field (MRF)
  - Each node x_i is “wet” or “dry”
  - Observations: Precipitation
  - Smoothness in space and time
  - Most likely state assignments
    - Each (lat,long,time) gets “wet” or “dry”
  - Advanced analysis
    - Soil moisture, hydrology/watershed models
    - Multiple states based on severity, e.g., lower quantiles



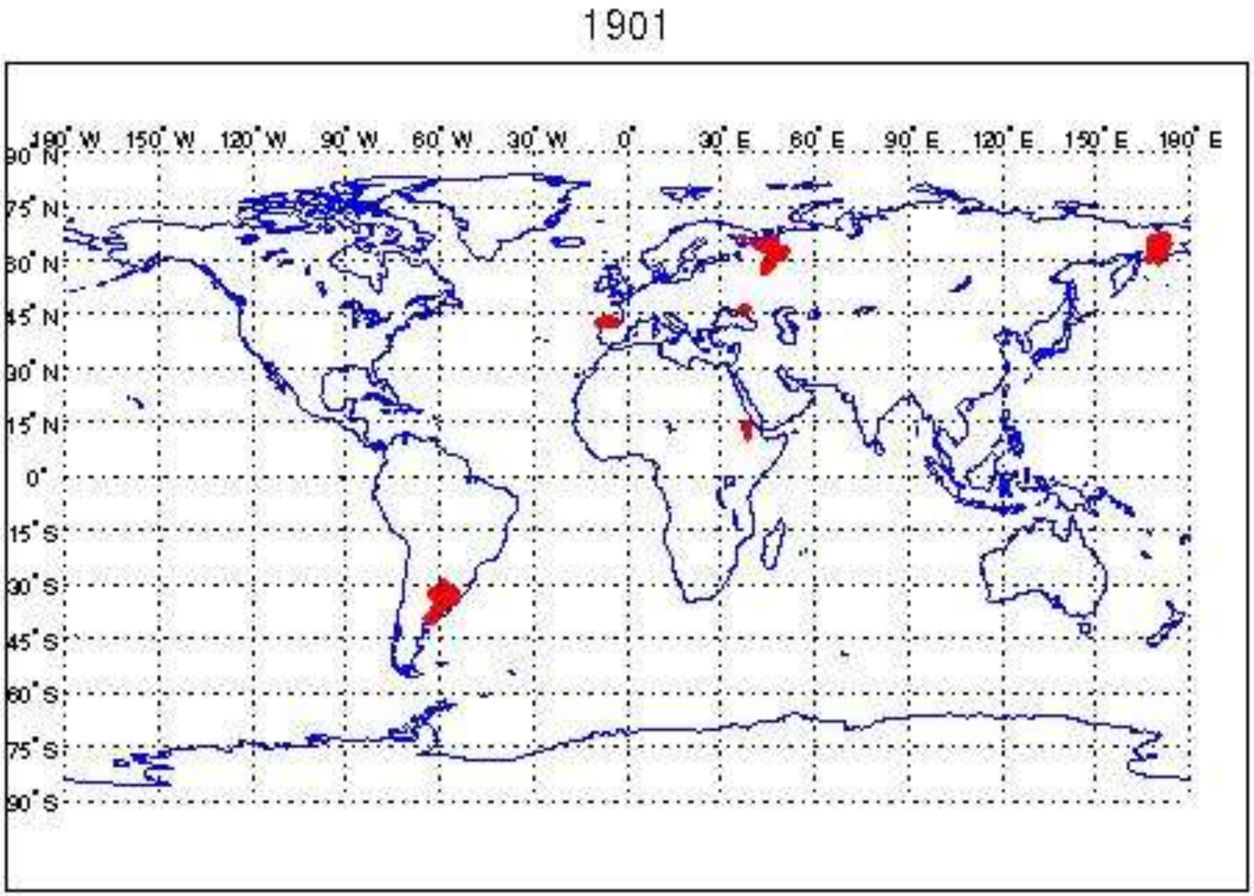
Results: Droughts starting in 1920-30s



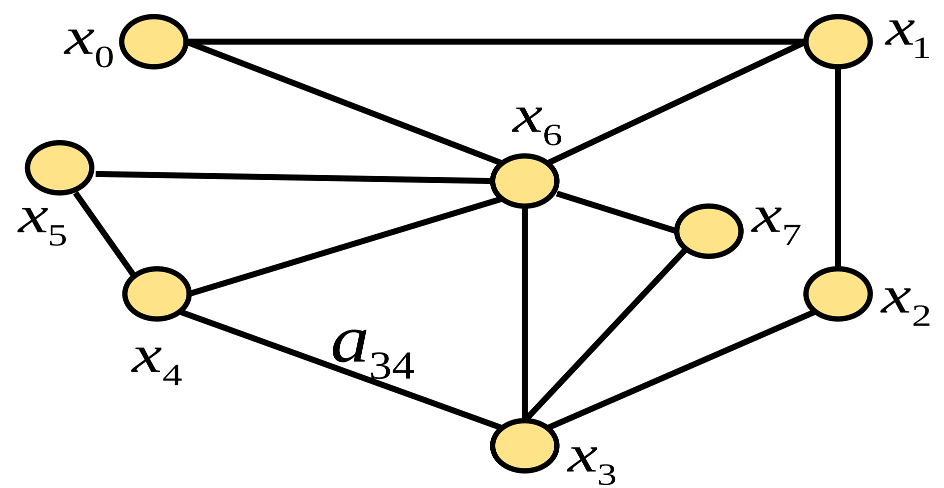
Results: Droughts starting in 1960-70s

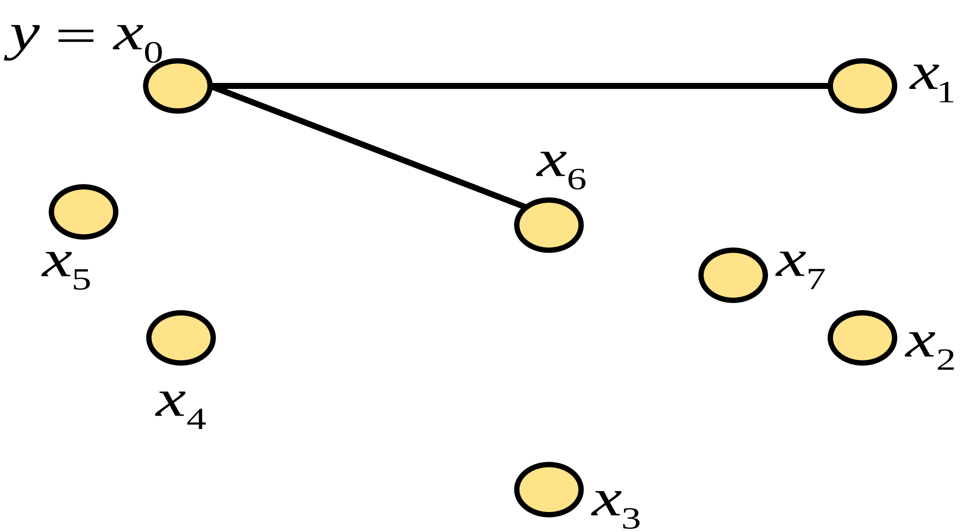


Major Droughts: 1901-2006



Learning dependencies





Dependencies in graphical models

- $a_{ij} = 0 \Leftrightarrow x_i \perp x_j \mid x_{-i,-j}$
- Example: $x_0 \mid x_5 \perp x_1, x_2, x_3, x_4, x_6, x_7$
- Conditional independence

Gaussian model $x \sim N(0,\Sigma)$

- Precision $A = \Sigma^{-1}$ is sparse

$$\mathrm{p}(\mathrm{x}|0, \mathrm{A}^{-1}) \propto \exp\left(-\sum_{i,j} a_{ij}\, xi\, xj\right)$$

- $a_{ij} = 0 \Leftrightarrow x_i \perp xj \mid x_{-i,-j}$

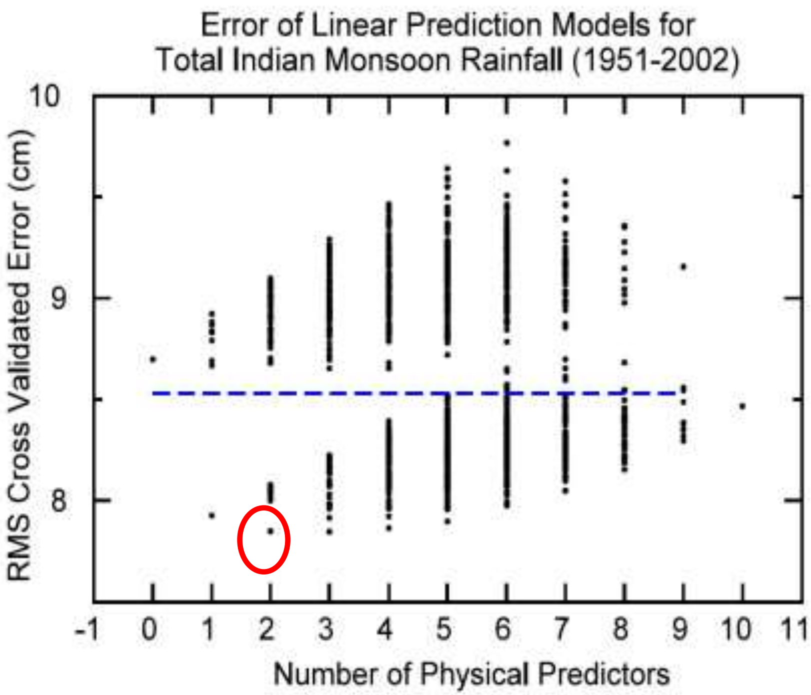
Estimating dependency structure

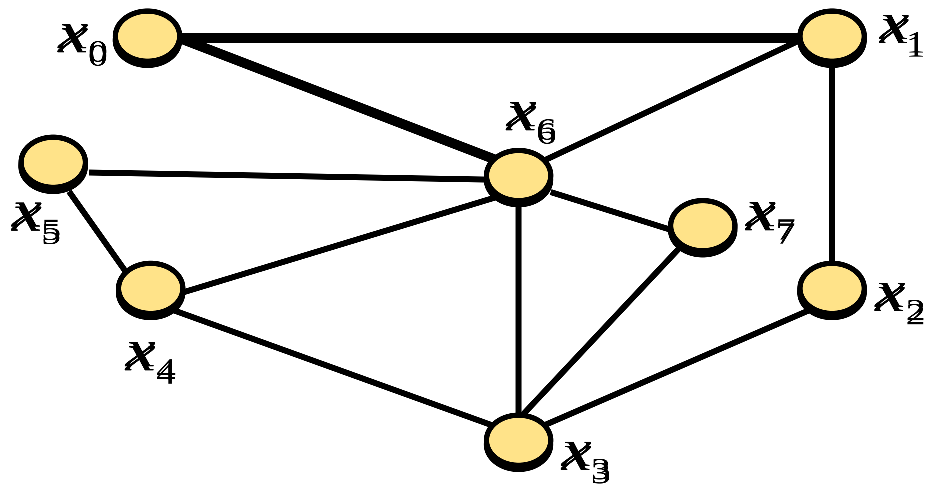
- One-vs-rest 'sparse regression'
- Lasso for multivariate Gaussians

$$\sum_{h=1}^{n} (y_h - \beta^T x_h)^2 + \lambda_n |\beta|_1$$

$$y = x_0, \qquad z = (x_1, \dots, x_7)$$

Sparse Regression, Structure Learning





Source: Delsole et al., 2002, 2006, 2009

Gaussian Copula: $(f_1(x_1),\ldots,f_i(x_i),\ldots,f_p(x_p)) \sim N(0, \Sigma)$, precision $A = \Sigma^{-1}$

- <u>Not</u> $(x_1,\ldots,x_i,\ldots,x_p) \sim N(0, \Sigma)$, f_i are monotonic transformations
- Sparse A can be consistently estimated (H. Liu et al., 2012)
- Linear Programming (LP) based estimator (CLIME) (T. Cai et al., 2011)
- LP estimator scales to high-dimensional copulas (Our work)
  - Millions of variables, trillions of edges

References

- K. Subbian and A. Banerjee, Climate Multi-model Regression Using Spatial Smoothing, *SIAM Data Mining (SDM)*, 2013. [Best Application Paper Award]

- Q. Fu, A. Banerjee, S. Liess, and P. Snyder. Drought Detection for the Last Century: A MRF-based Approach, *SIAM Data Mining (SDM)*, 2012.
- Q. Fu, H. Wang, and A. Banerjee, Bethe-ADMM for Tree Decomposition based Parallel MAP inference, *Uncertainty in Artificial Intelligence (UAI)*, 2013.
- C. Jin, Q. Fu, H. Wang, A. Agrawal, W. Hendrix, W.-K. Liao, M. A. Patwary, A. Banerjee, and A. Choudhary, Solving Combinatorial Optimization Problems using Relaxed Linear Programming: A High Performance Computing Perspective, *BigMine workshop (KDD)*, 2013. [Best Paper Award]

- C. Hsieh, I. Dhillon, P. Ravikumar, A. Banerjee, A Divide-and-Conquer Method for Sparse Inverse Covariance Estimation, *Advances in Neural Information Processing Systems (NIPS)*, 2012.
- S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, Sparse Group Lasso: Consistency and Climate Applications, *SIAM Data Mining (SDM)*, 2012. [Best Student Paper Award]