Graphical Models for Climate Data Analysis: Drought Detection and Land Variable Regression

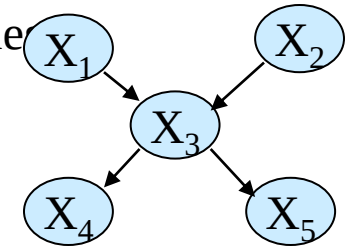
Arindam Banerjee
banerjee@cs.umn.edu
Dept of Computer Science & Engineering
University of Minnesota, Twin Cities

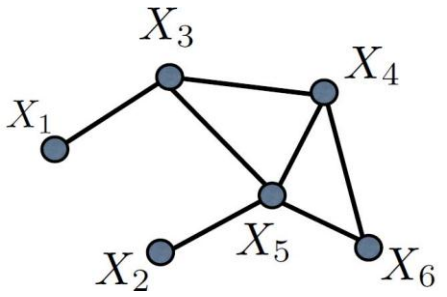
Joint Work with:
Snigdhansu Chatterjee, Soumyadeep Chatterjee,
Auroop Ganguly, Stefan Liess, Peter Snyder

First Workshop on Understanding Climate Change from Data
August 15-16, 2011

Graphical Models: What and Why

- Graphical models
  - ependencies between (random) variables
  - Closer to reality, learning/inference is much more difficult
- Basic nomenclature
  - Node = Random Variable (hidden/observed), Edge = Statistical Dependency
- Directed Graphs
  - A *directed* graph between random variables, causal dependencie
  - 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 for Climate Data Analysis

- Abrupt Change Detection
  - Climate change is not always smooth and gradual
  - Sudden and Large changes in environmental conditions
  - First Step: Detecting Abrupt Changes
  - Focus: Drought Detection

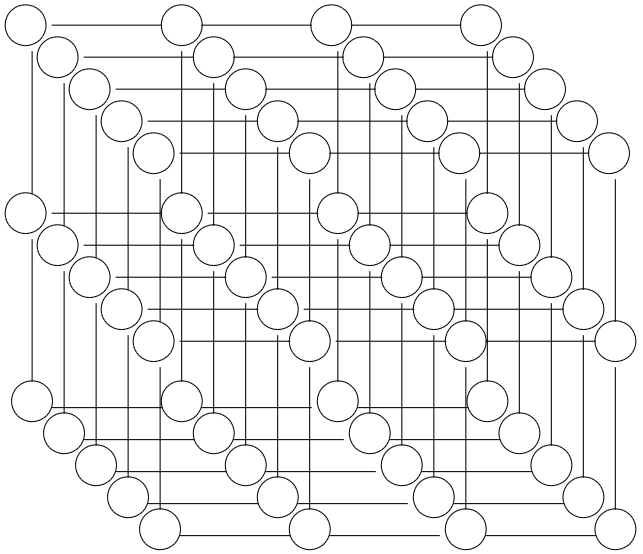
- Predictive Modeling
  - High-dimensional problems, Nonlinear dependencies
  - The “High p, Low n” regime
  - First Step: Effective high-dimensional regression for climate problems
  - Focus: Land variable regression

Drought Detection

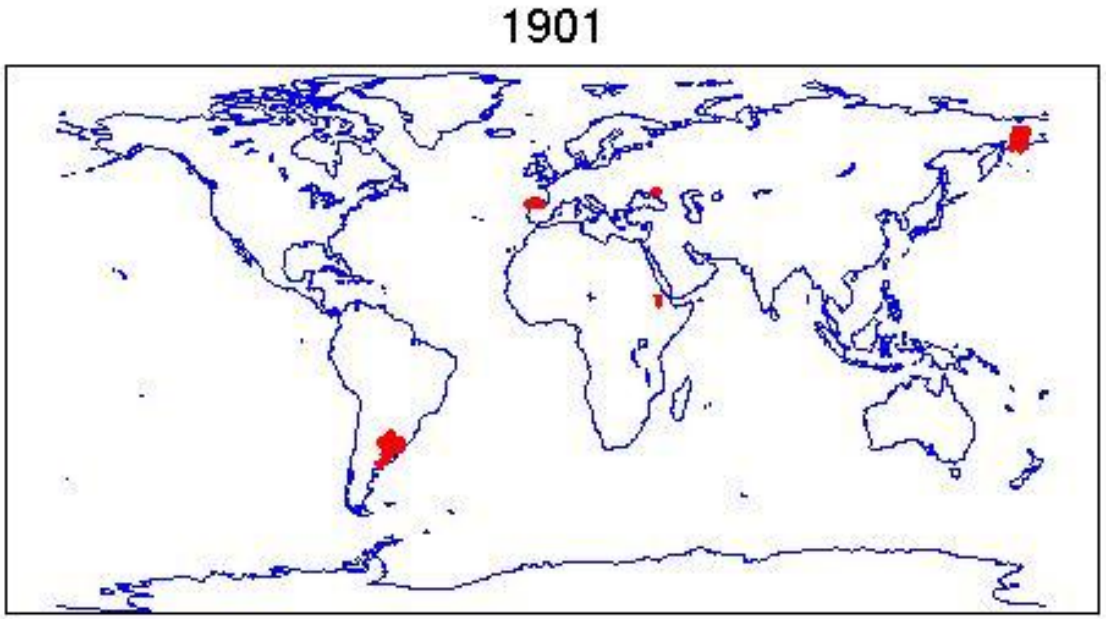
- Significant droughts
  - Persistent over space and time
  - Catastrophic consequences
- Examples:
  - Late 1960s Sahel drought
  - 1930s North American Dust Bowl
- Dataset: Climate Research Unit
  - http://data.giss.nasa.gov/precip_cru/
  - Monthly precipitation over land
  - Duration: 1901 to 2006
  - Resolution: 0.5 latitude X 0.5 longitude

Detection with Markov Random Field (MRF)

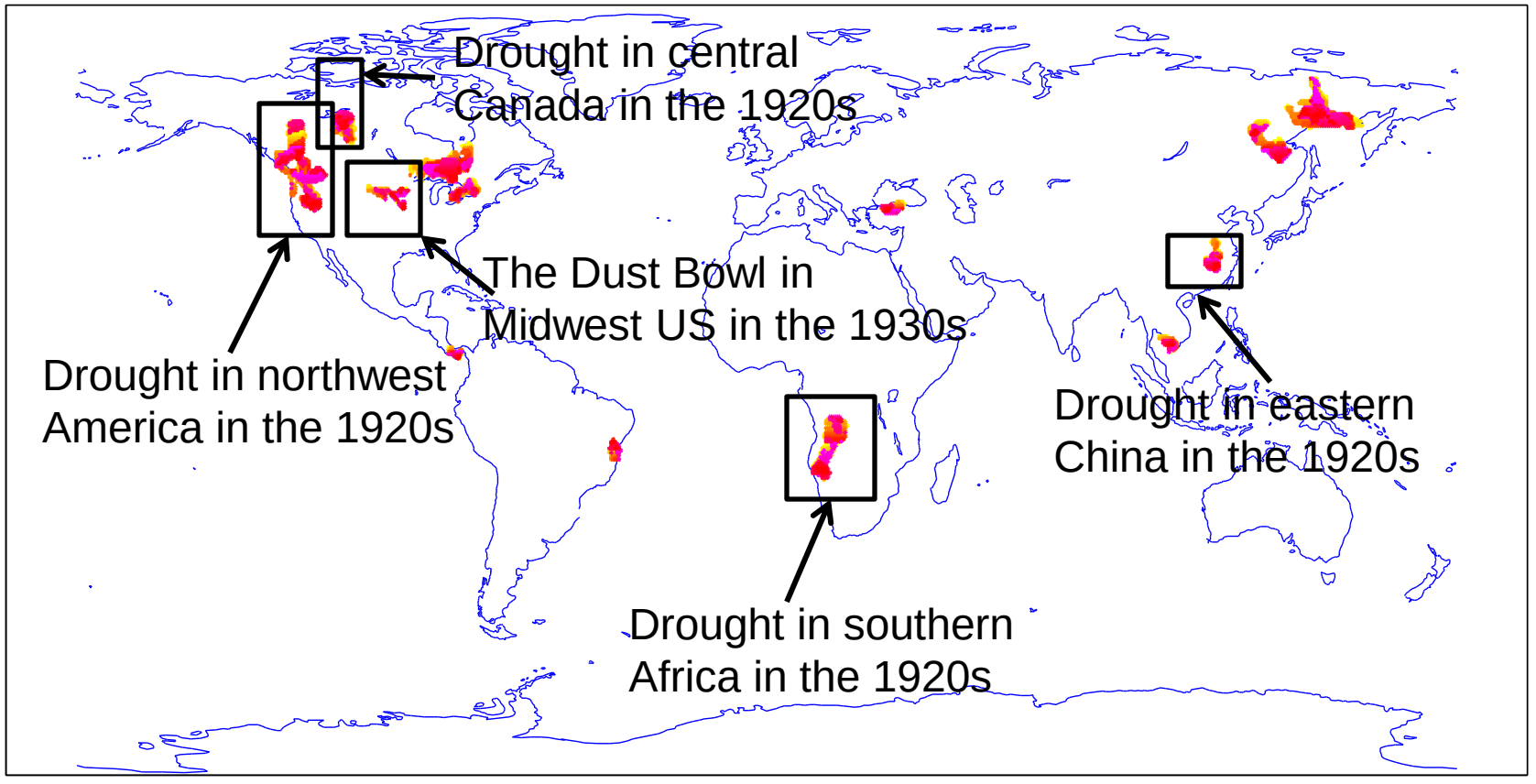
- Model dependencies using a 4-nearest neighbor grid
  - Replicate grid over time
  - Each node can be 0 (normal) or 1 (dry)
- Toy example:
  - m = 3, n = 4, N = 12, T = 5
  - Total # States: $2^{60} = 1.1529 \times 10^{18}$
- CRU data:
  - m = 720, n = 360, N = 67420, T = 106
  - Total # States: $2^{7146520} > 10^{2382200}$
- MAP Inference:
  - Find the most likely "state" of the system
  - Integer programming with LARGE number of states
  - Postprocess MAP state to identify significant droughts



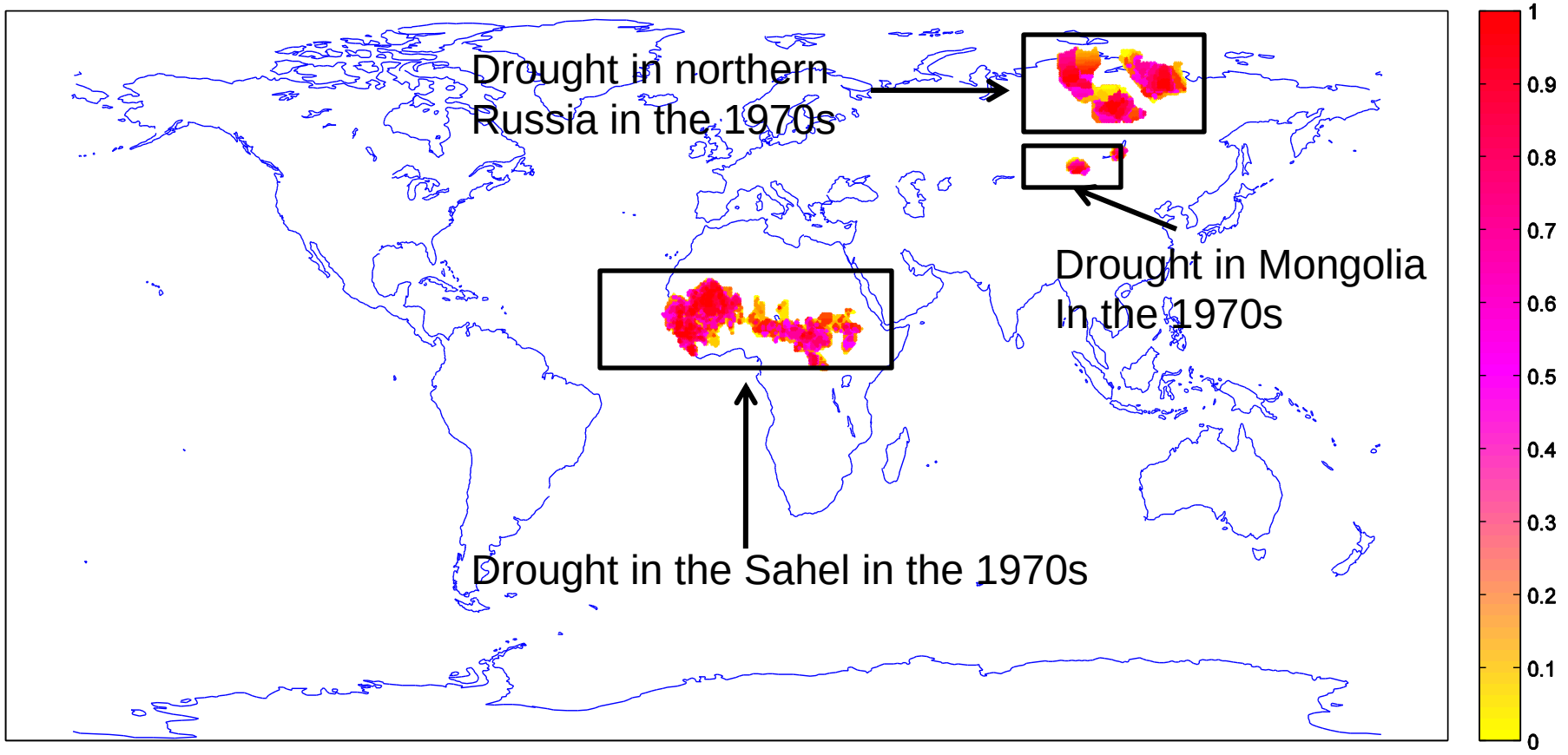
Drought Regions from MRFS



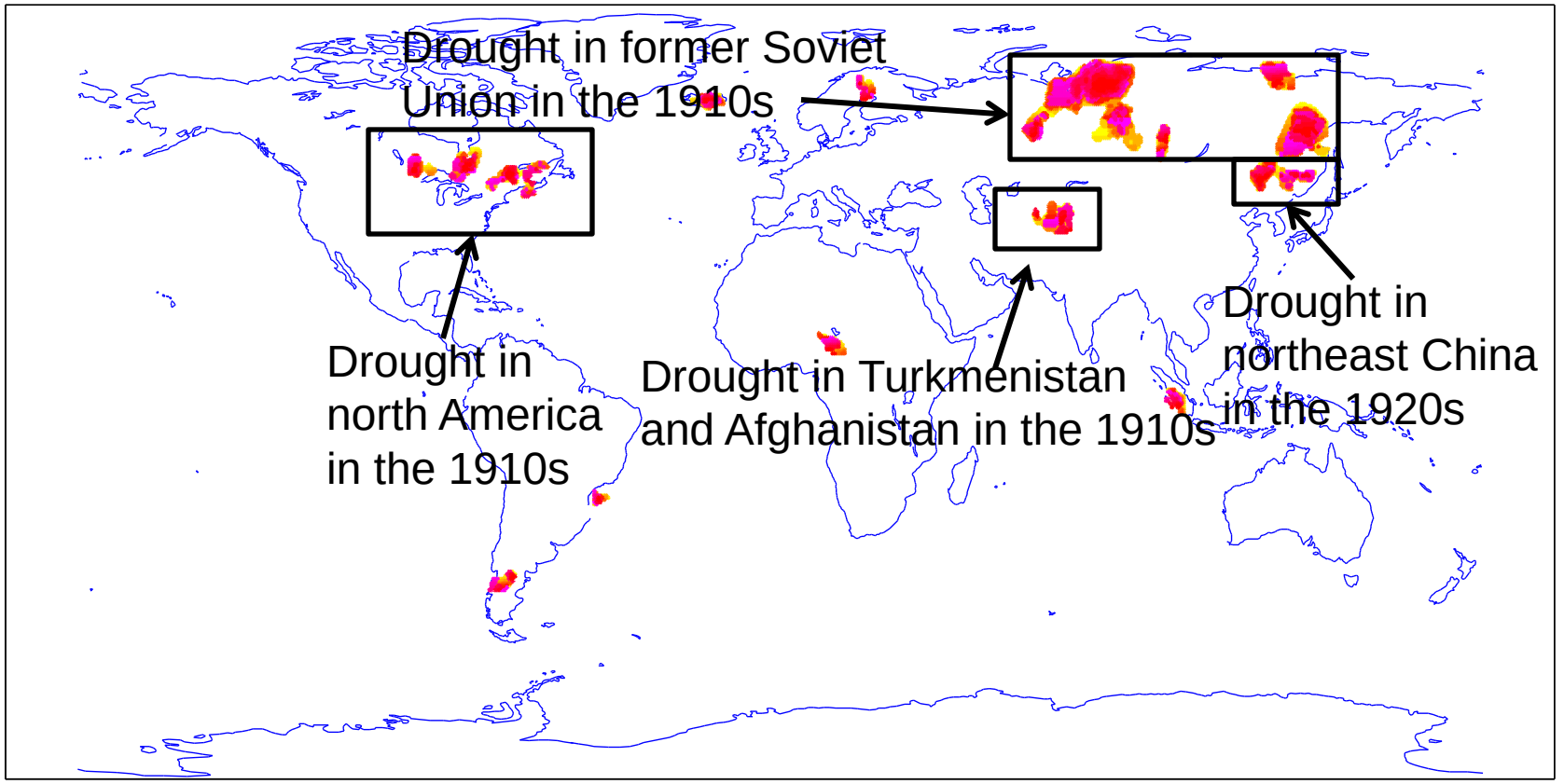
Results: Drought Detection



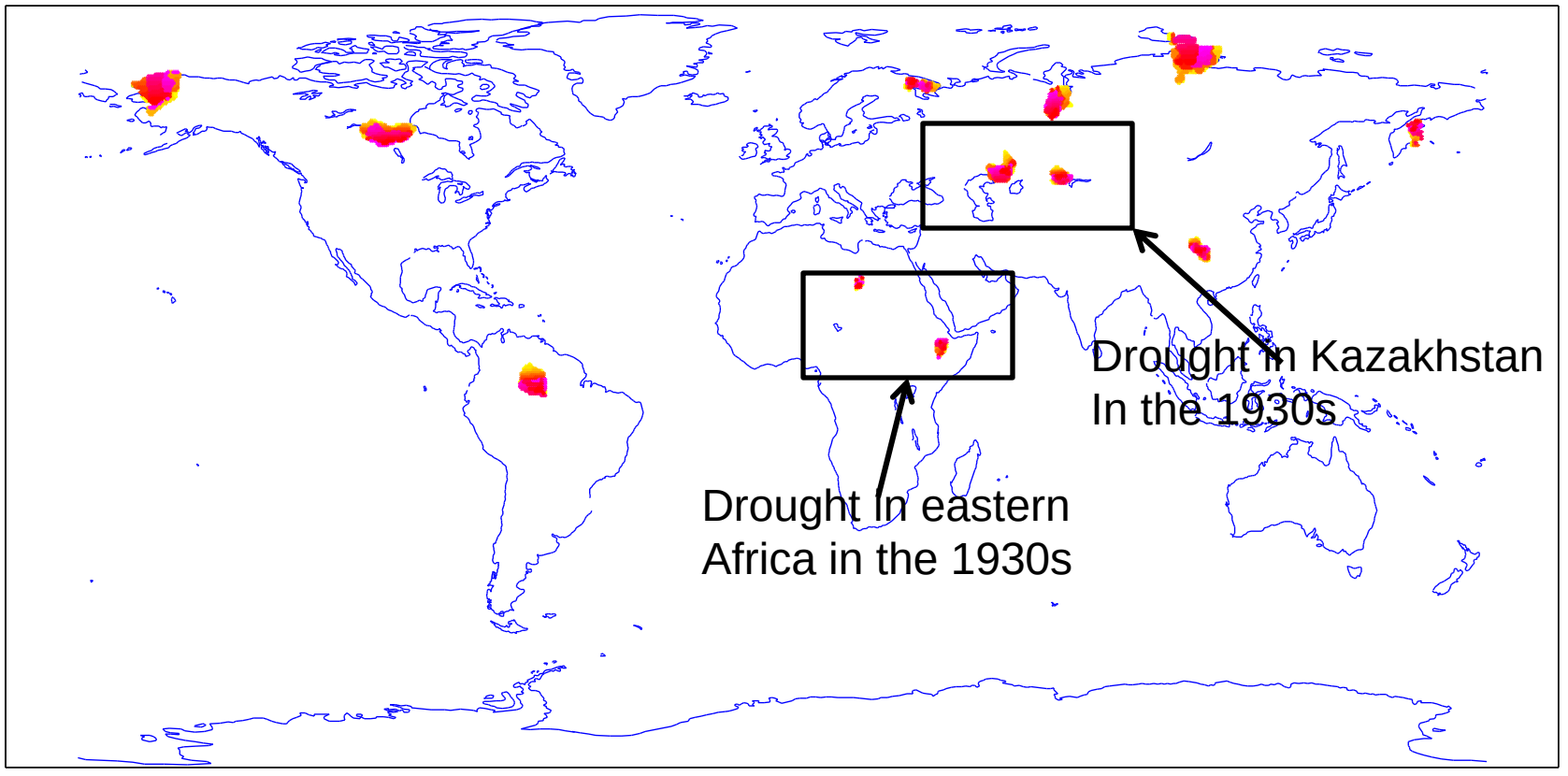
Results: Drought Detection



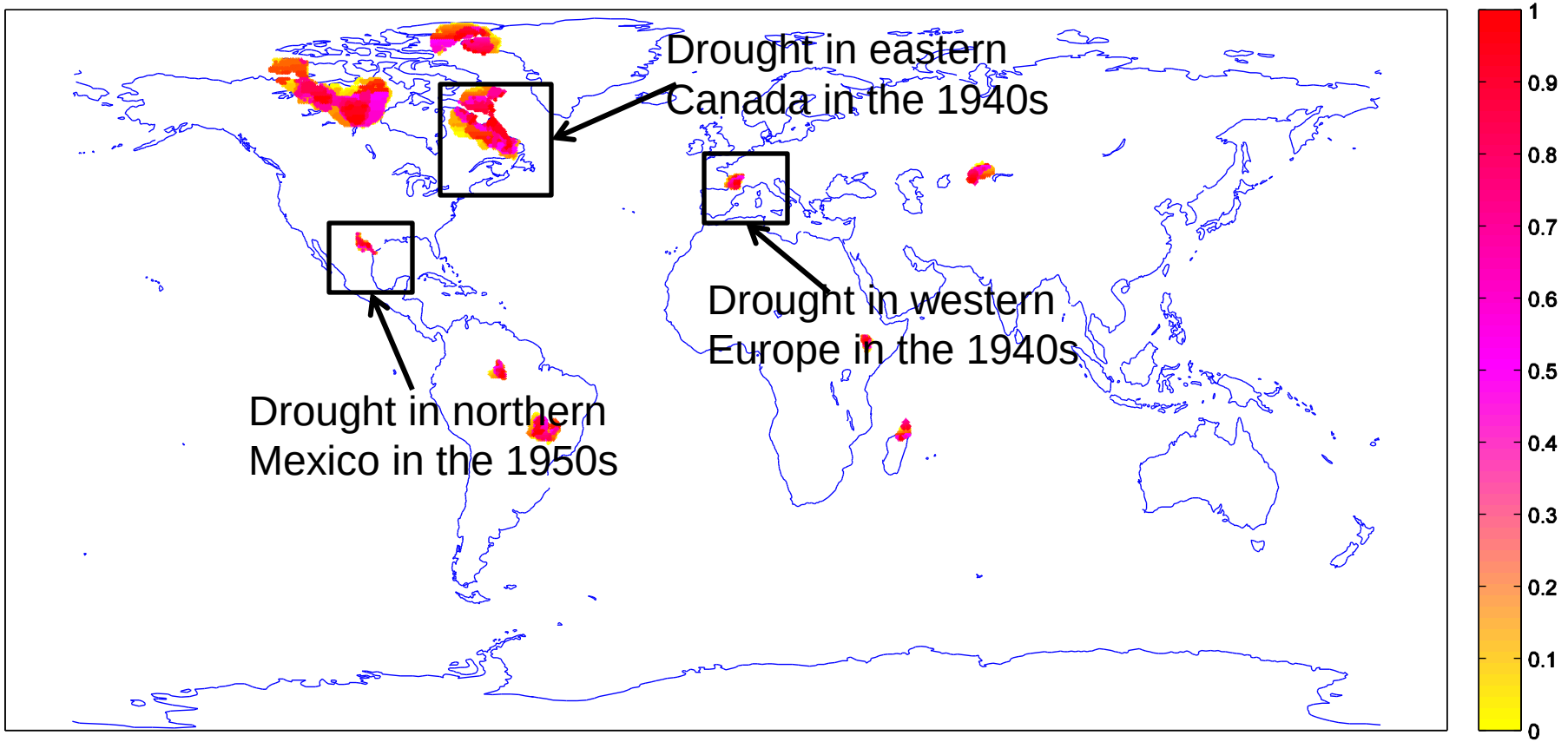
Results: Drought Detection



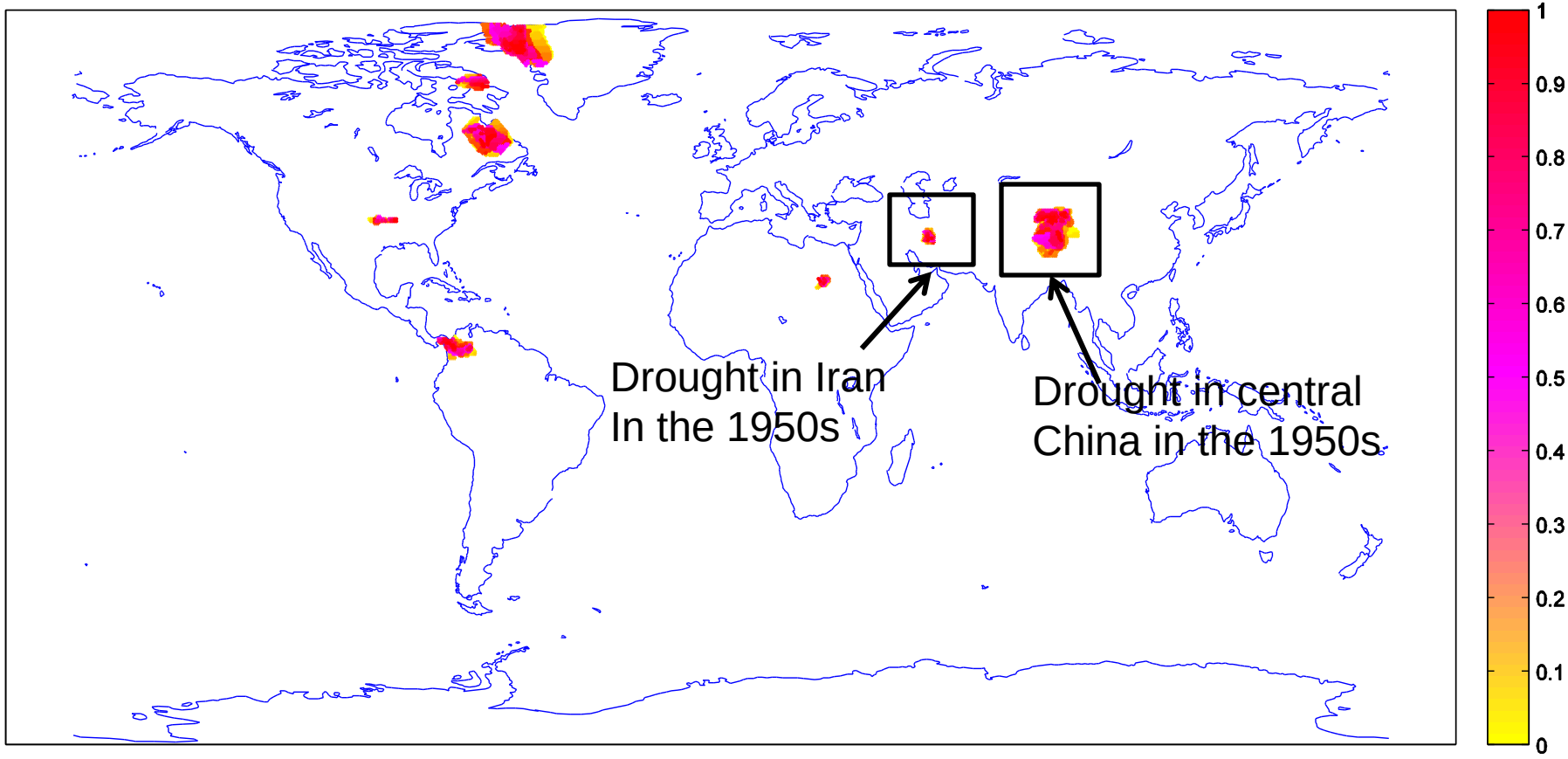
Results: Drought Detection



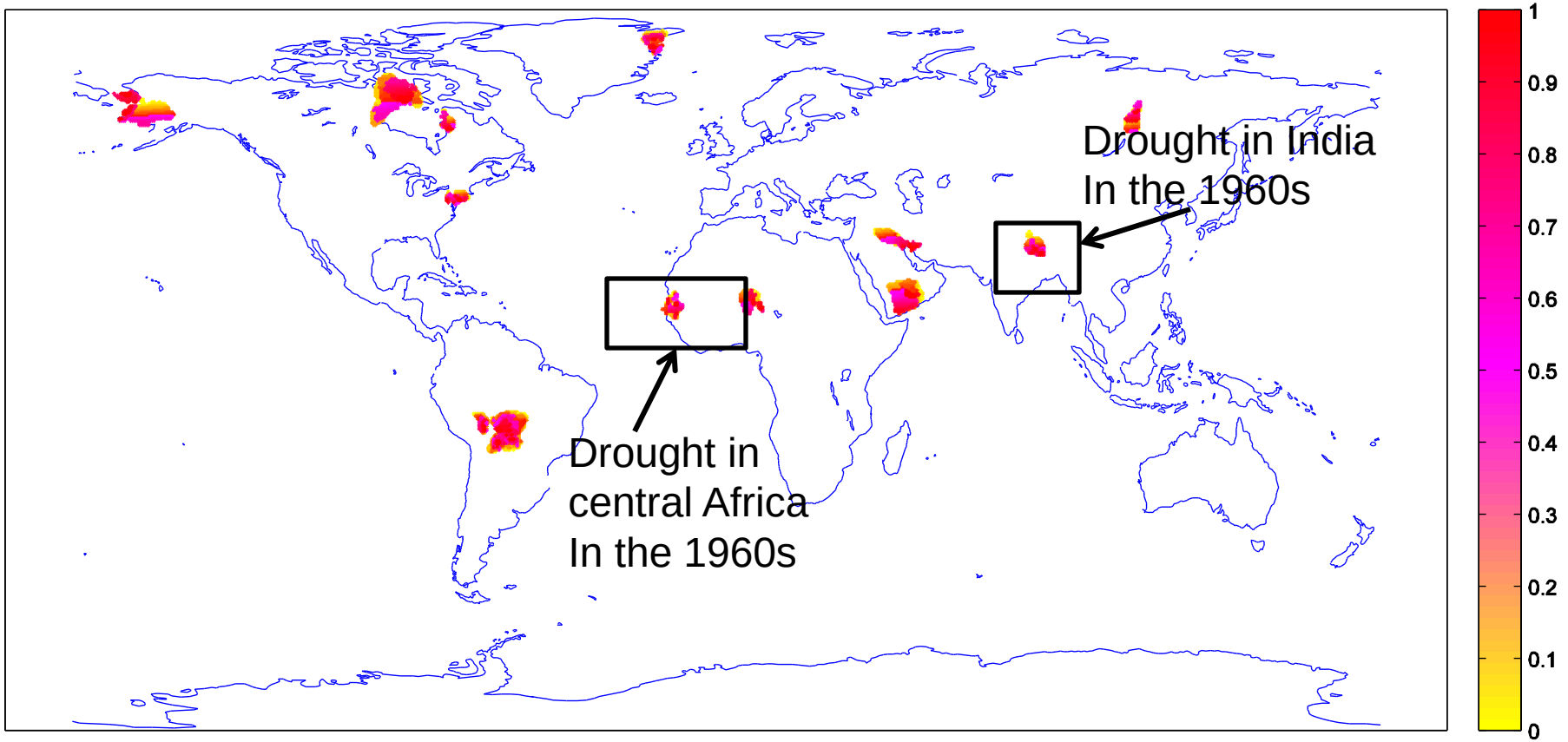
Results: Drought Detection



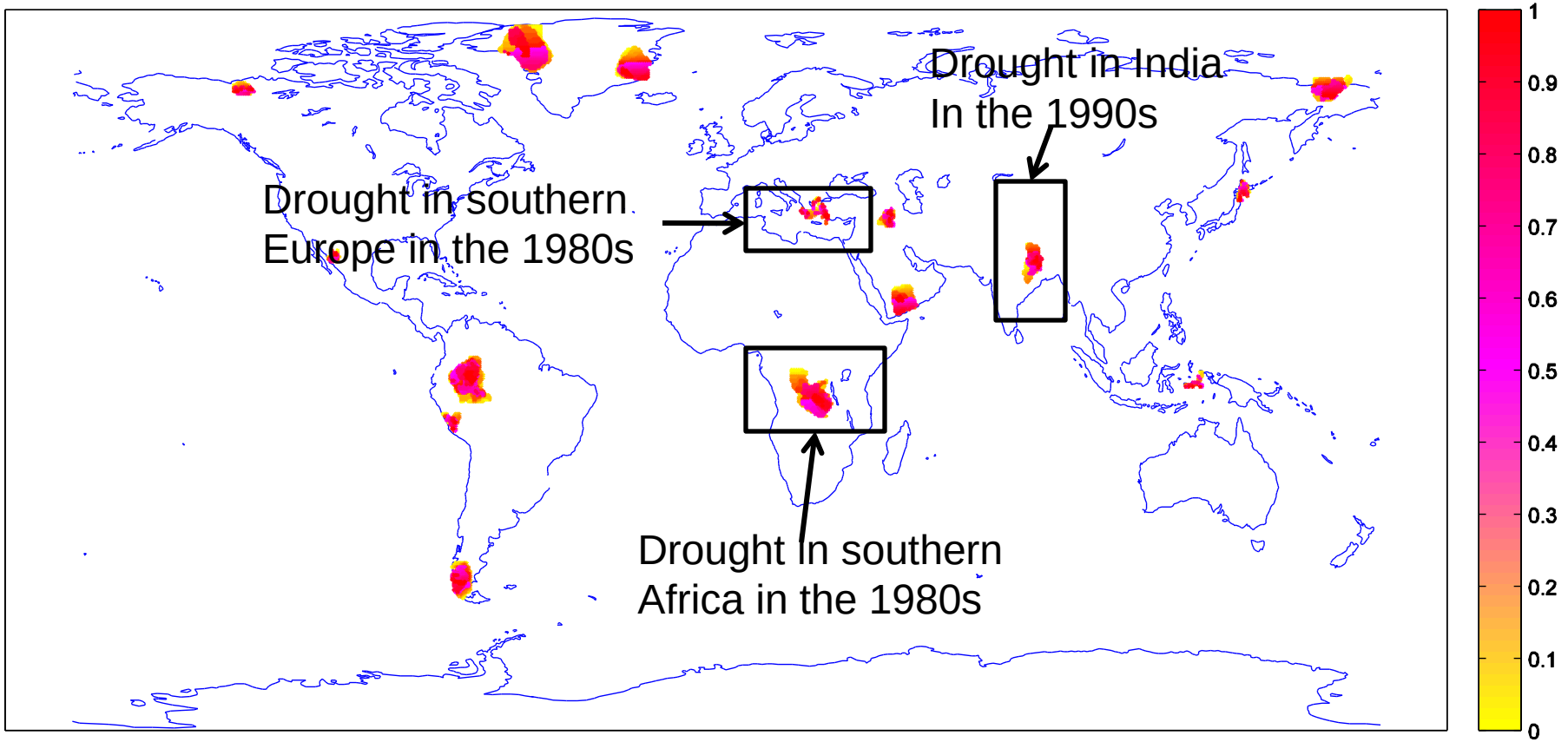
Results: Drought Detection



Results: Drought Detection



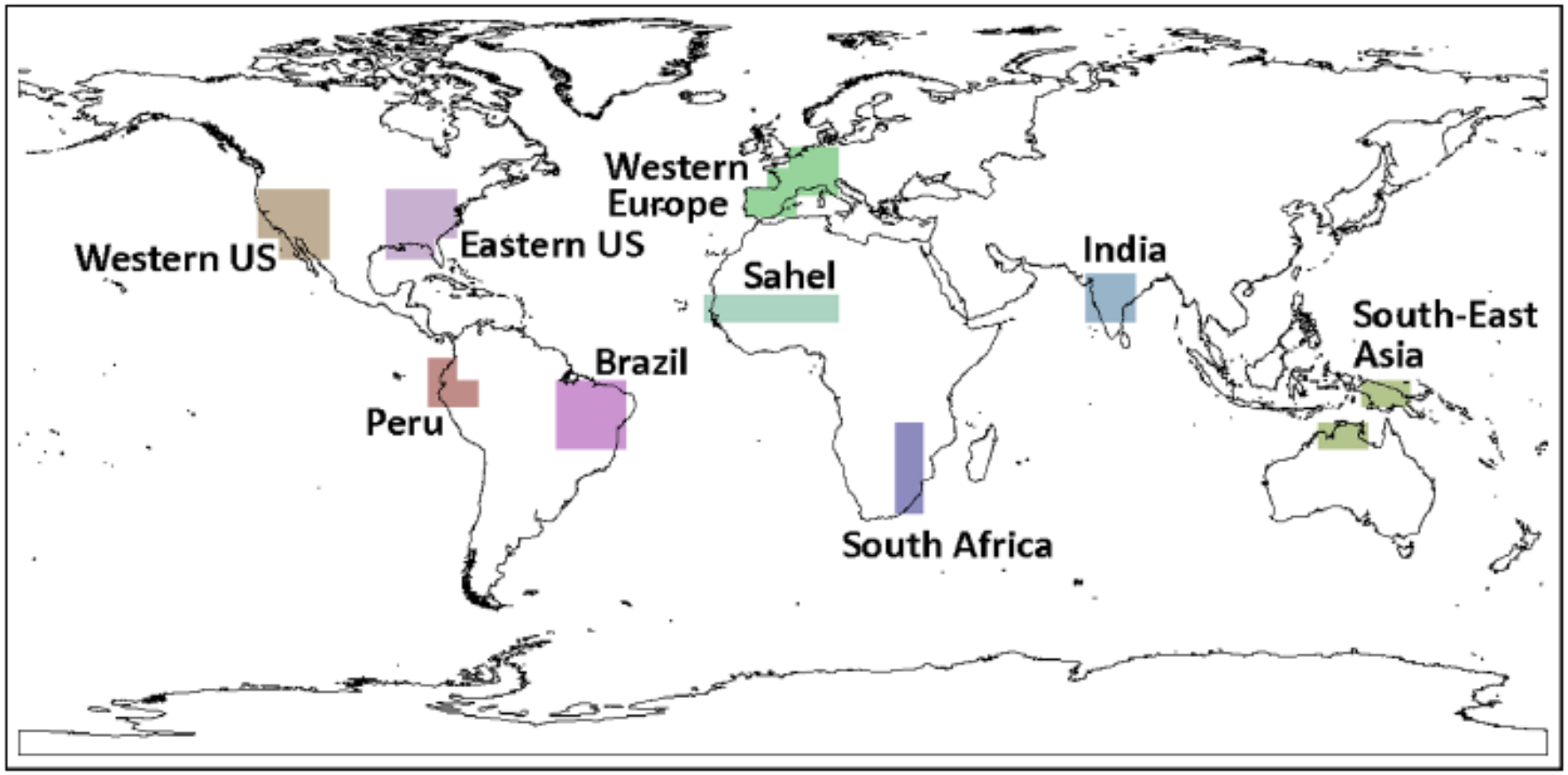
Results: Drought Detection



Land Variable Regression

- Land-Sea variable interactions
  - How do sea variables affect land variables?
  - Are proximal locations important?
  - Are there long range spatial dependencies (tele-connections)?
- NCEP/NCAR Reanalysis 1: monthly means for 1948-2010
  - Covariates: Temperature, Sea Level Pressure, Precipitation, Relative Humidity, Horizontal Wind Speed Vertical Wind Speed
  - Response: Temperature, Precipitation at 9 locations
- High-dimensional Regression
  - Dimension $p = Lm$, L locations, m variables/location
  - Two considerations:
    - Not all locations are relevant
    - Even if a location is relevant, not all variables are relevant

Land Regions for Prediction: Temp, Precip



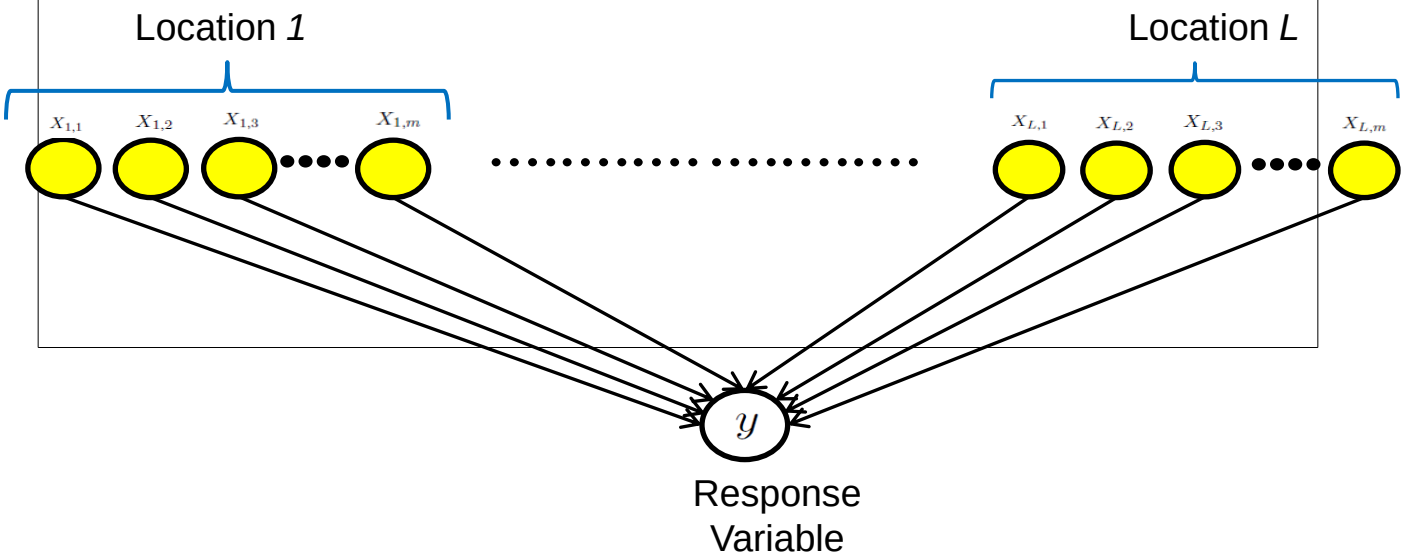
Land regions chosen for predictions

Ordinary Least Squares

- Naive method: Use *ordinary least squares* (OLS) :

$$\hat{\theta} := \arg min_{\theta \in \mathbb{R}^{mL}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\}$$

- For $n < mL$, OLS estimate $\hat{\theta}$ non-unique
- In general, y depends on all mL covariates: complex model



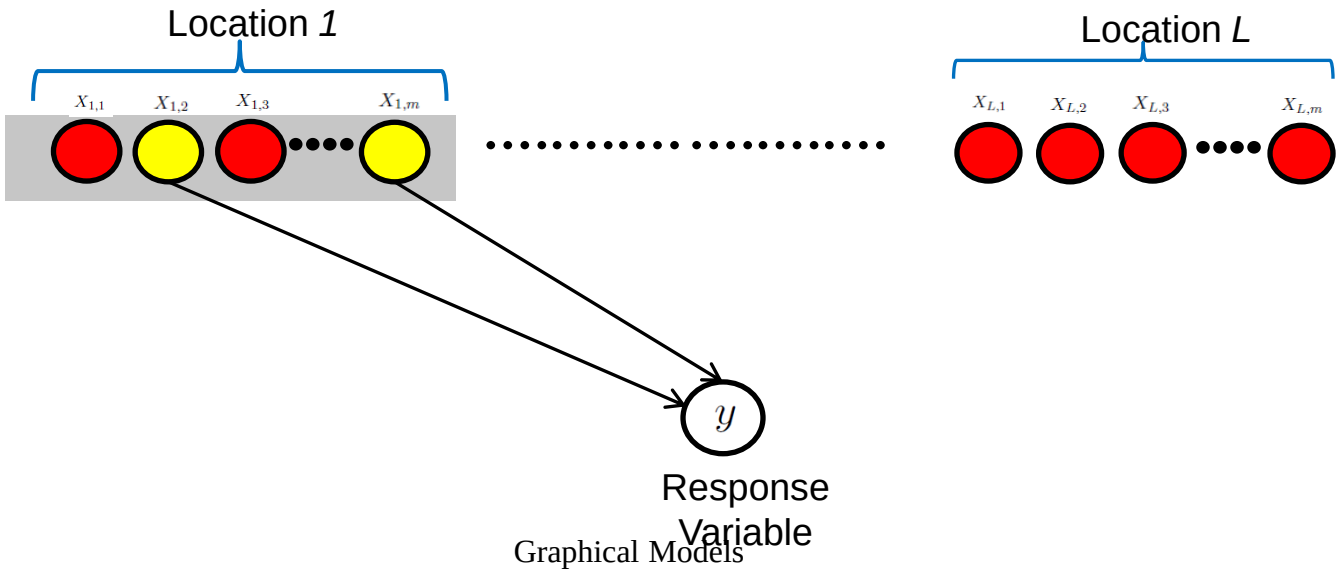
Sparse Group Lasso (SGL)

- High-dimensional Regression with "Sparse" "Group Sparsity"

$$\hat{\theta}_{SGL} = \arg min_{\theta \in \mathbb{R}^{mL}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_{1,\mathcal{G}} \right\}$$

$$\|\theta\|_1 = \sum_{i=1}^{mL} |\theta_i| \qquad \|\theta\|_{1,\mathcal{G}} = \sum_{k=1}^{L} \|\theta_{G_k}\|_2$$

$$\mathcal{G} = \{G_1, \ldots, G_L\}: \text{ groups of } m \text{ variables at } L \text{ locations}$$



Sparse Group Lasso

- Optimization methods:
  - Proximal method: Block-coordinate Dual Ascent
  - Gradient based Primal Method
- Consistency and Rates of Convergence:

Theorem *Let A be any subspace of $\mathbb{R}^p$ of dimension s_A. Let θ^* be the optimal (unknown) regression parameter, and let $r^{A^\perp}_{(1,\mathcal{G}_\nu,\alpha)}(\theta^*)$ be the sparse-group lasso norm of θ^* restricted to $A^\perp$, the orthogonal subspace of A. Then, if λ_n satisfies the lower bound in Lemma 1, with probability at least $(1 - \frac{2}{(pT)^k})$, we have*

$$\|\hat{\theta_{\lambda_n}} - \theta^*\|_2^2 \leq \frac{4\lambda_n^2}{k_{\mathcal{L}}^2} s_A + \frac{2\lambda_n}{k_{\mathcal{L}}} r^{A^\perp}_{(1,\mathcal{G}_\nu,\alpha)}(\theta^*) ,$$

where $\hat{\theta_{\lambda_n}}$ is the SGL estimator

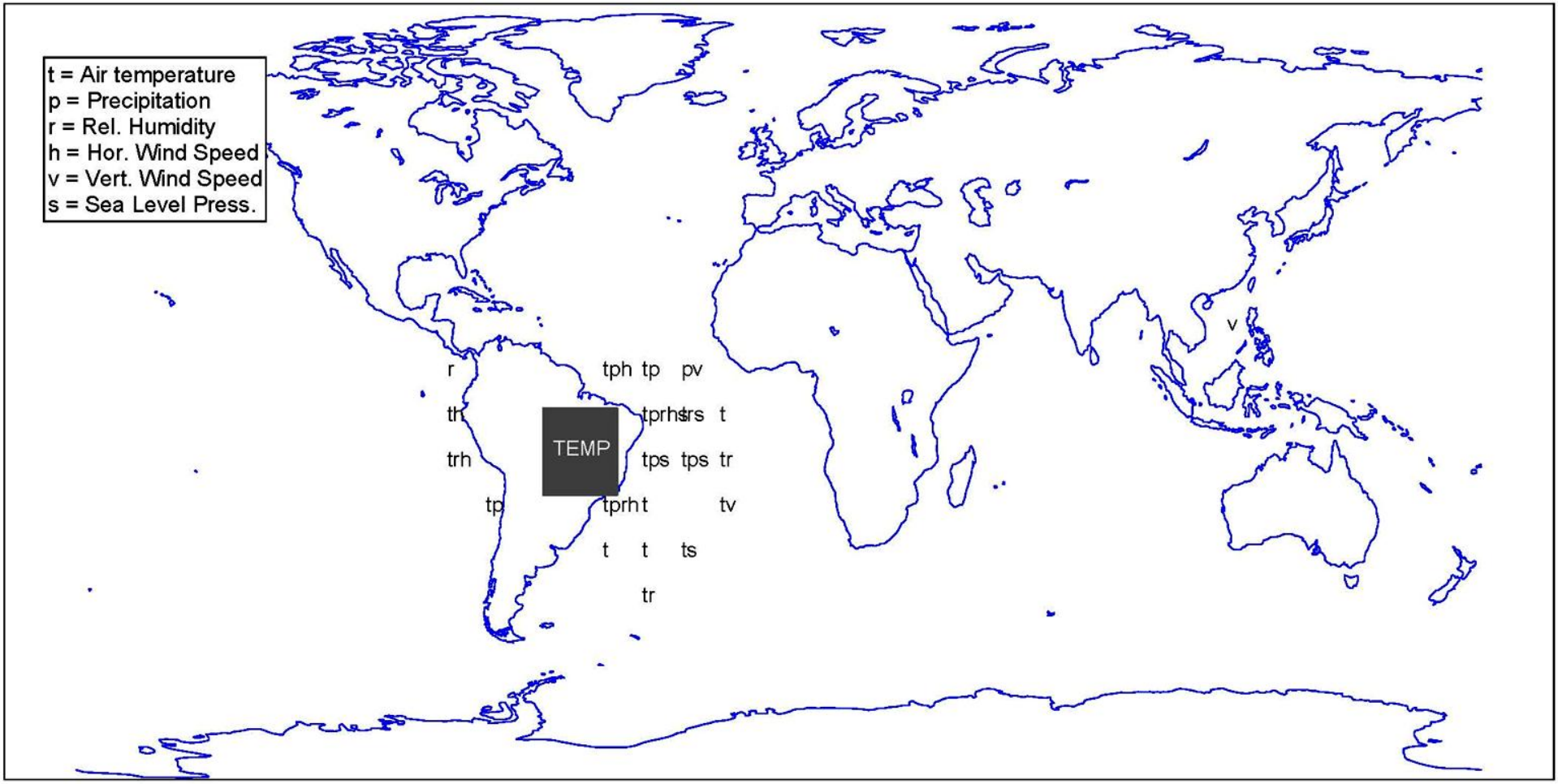
Corollary *If the optimal parameter θ^* is in the subspace A, then*

$$\|\hat{\theta_{\lambda_n}} - \theta^*\|_2^2 \leq \frac{4\lambda_n^2}{k_{\mathcal{L}}^2} s_A = O\left(\frac{\log p}{n}\right)$$

Error (RMSE): SGL, NC, OLS

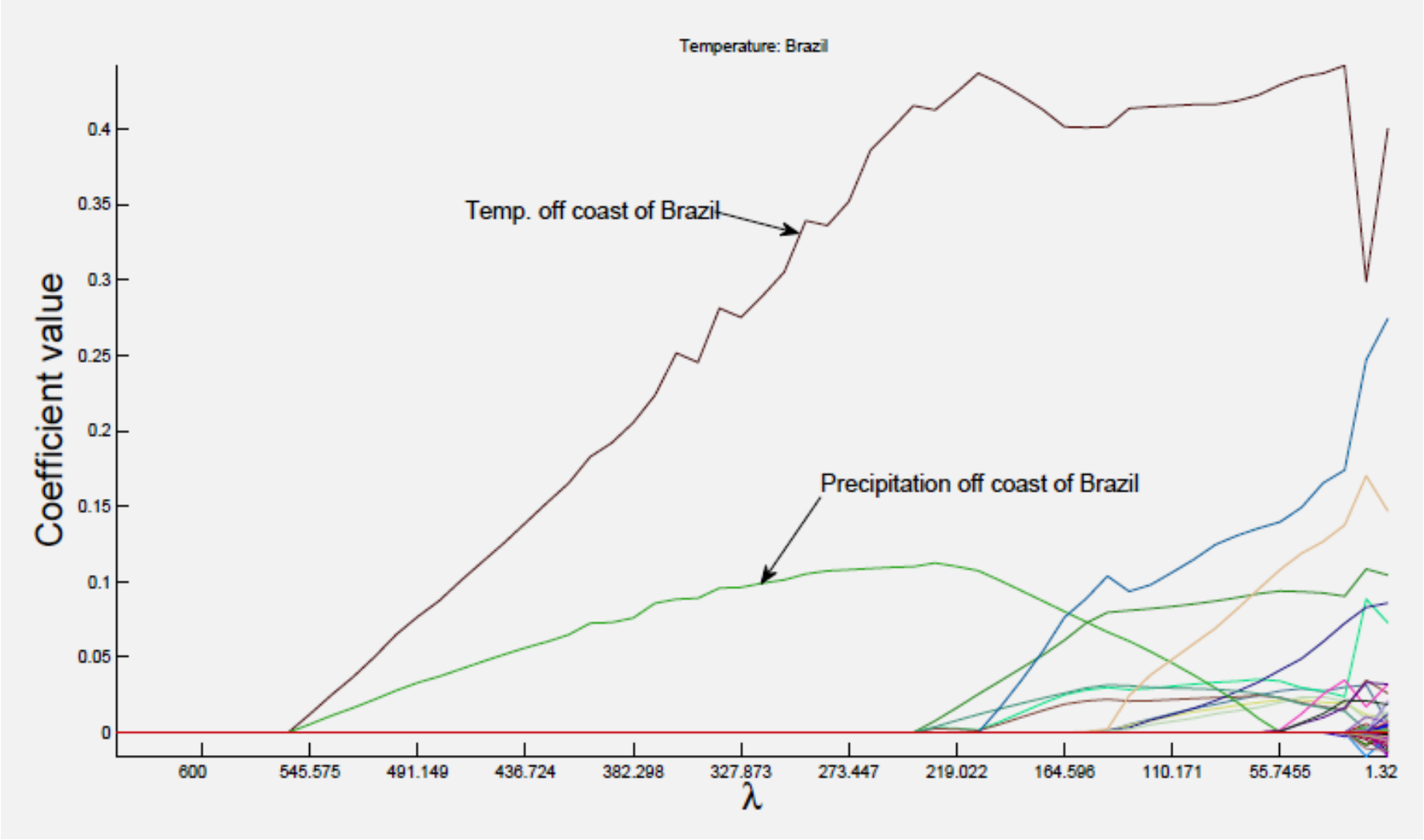
Variable	Region	SGL	Network Clusters	OLS
Air Temperature	Brazil	**0.198**	0.534	0.348
	Peru	**0.247**	0.468	0.387
	West USA	**0.270**	0.767	0.402
	East USA	**0.304**	0.815	0.348
	W Europe	**0.379**	0.936	0.493
	Sahel	**0.320**	0.685	0.413
	S Africa	**0.136**	0.726	0.267
	India	**0.205**	0.649	0.300
	SE Asia	**0.298**	0.541	0.383
Precipitation	Brazil	**0.261**	0.509	0.413
	Peru	**0.312**	0.864	0.523
	West USA	**0.451**	0.605	0.549
	East USA	**0.365**	0.686	0.413
	W Europe	**0.358**	0.45	0.551
	Sahel	**0.427**	0.533	0.523
	S Africa	**0.235**	0.697	0.378
	India	**0.146**	0.672	0.264
	SE Asia	**0.159**	0.665	0.312

SGL: Regression with Feature Selection



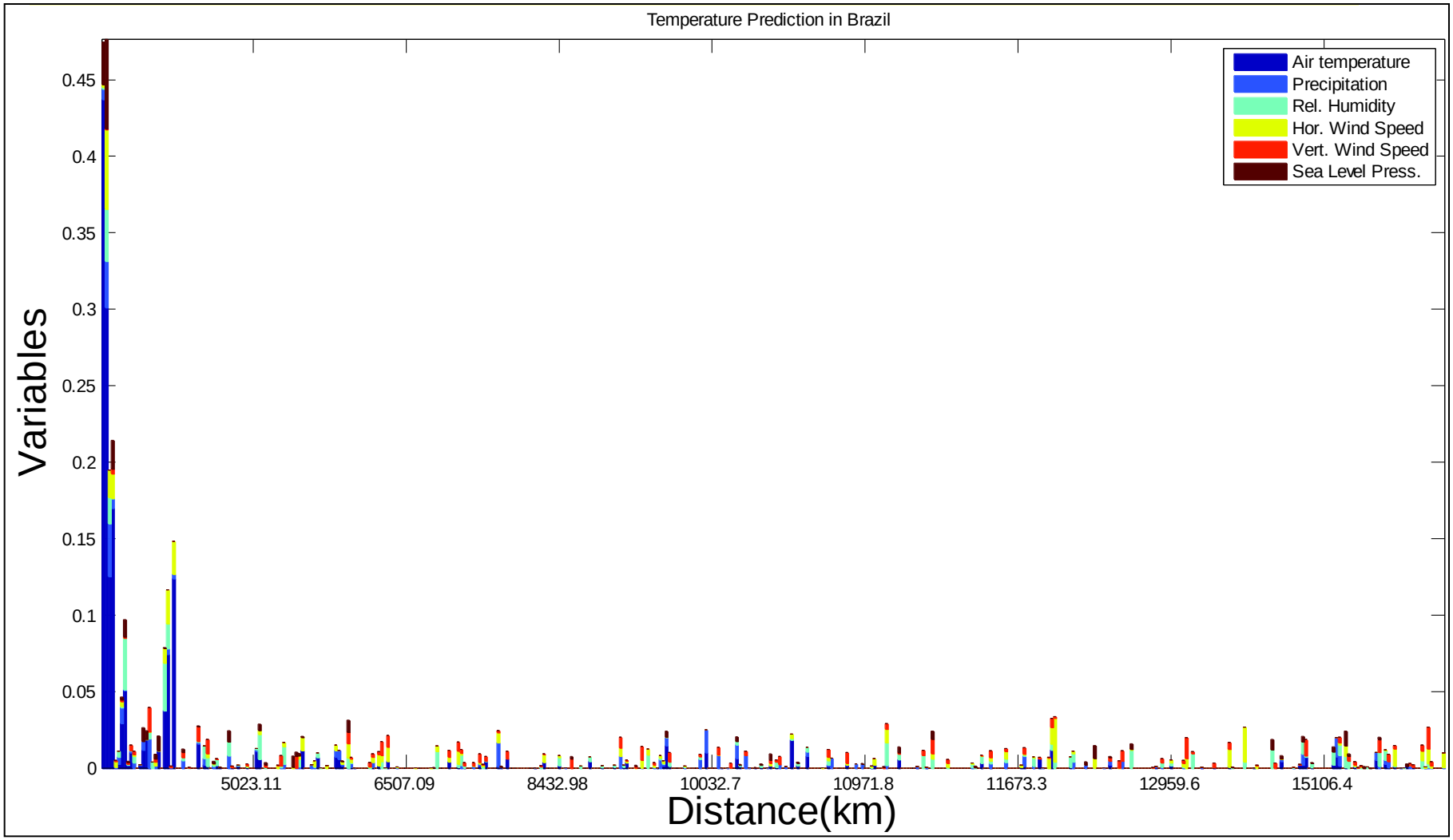
Relevant Variables for Temperature Prediction in Brazil

Robustness: Regularization Path of SGL



Regularization Path: Temperature prediction in Brazil

Spread of Coefficients for SGL



Component bar chart of coefficient magnitudes of variables for temperature prediction in Brazil

Summary

- Graphical Models for Climate Data Analysis
  - "Networks" over random variables
  - Captures dependencies among variables or "nodes"
- Abrupt Change Detection
  - Focus: Drought Detection using MRFs
  - Other univariate/multivariate abrupt change detection
- Predictive Modeling
  - Focus: Land variable regression using SGL
  - Other high-dimensional predictive models
    - Sparse Structure Learning
    - Sparse Extreme-value or Quantile Regression