Understanding Dominant Factors for Precipitation in Great Lakes Region

Soumyadeep Chatterjee

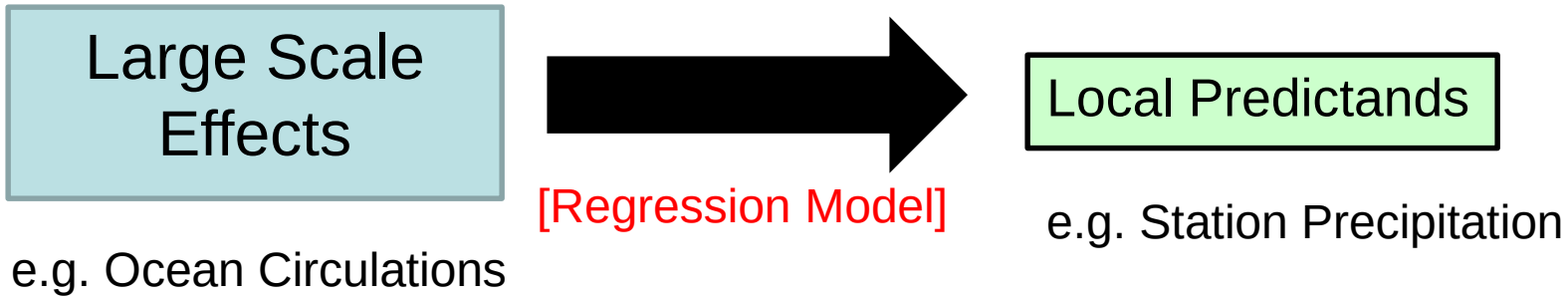
Dept. of Computer Science & Engineering
University of Minnesota, Twin Cities

Fifth Workshop on Understanding Climate Change from Data
August 4, 2015

Today

- Sparse Models for feature selection
- Problem: Factors affecting precipitation
- Finding Dominant Factors

Statistical Dependencies

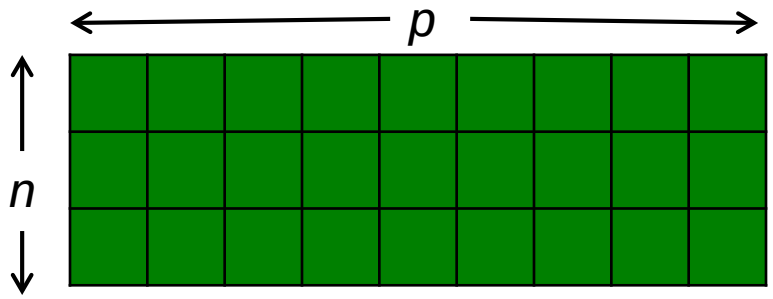


Key Requirements:

- Understanding of statistical dependence
- Feature Selection from many possible predictors
  - E.g. given by a domain expert, derived variables etc.

Feature Selection

- Regression in LOW dimensions ($p<n$)
- Challenges
  - Overfitting
  - Dependent covariates
  - Dependent samples
- Sparse Regression
  - Select dominant features
- Testing robustness of selected features
- Hypothesis generation for dominant factors

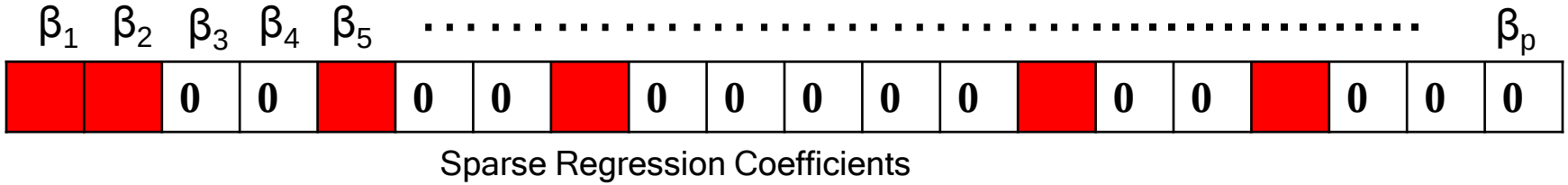


LASSO (Sparse Regression)

- Estimation under “structural constraints”, e.g. sparsity
- LASSO

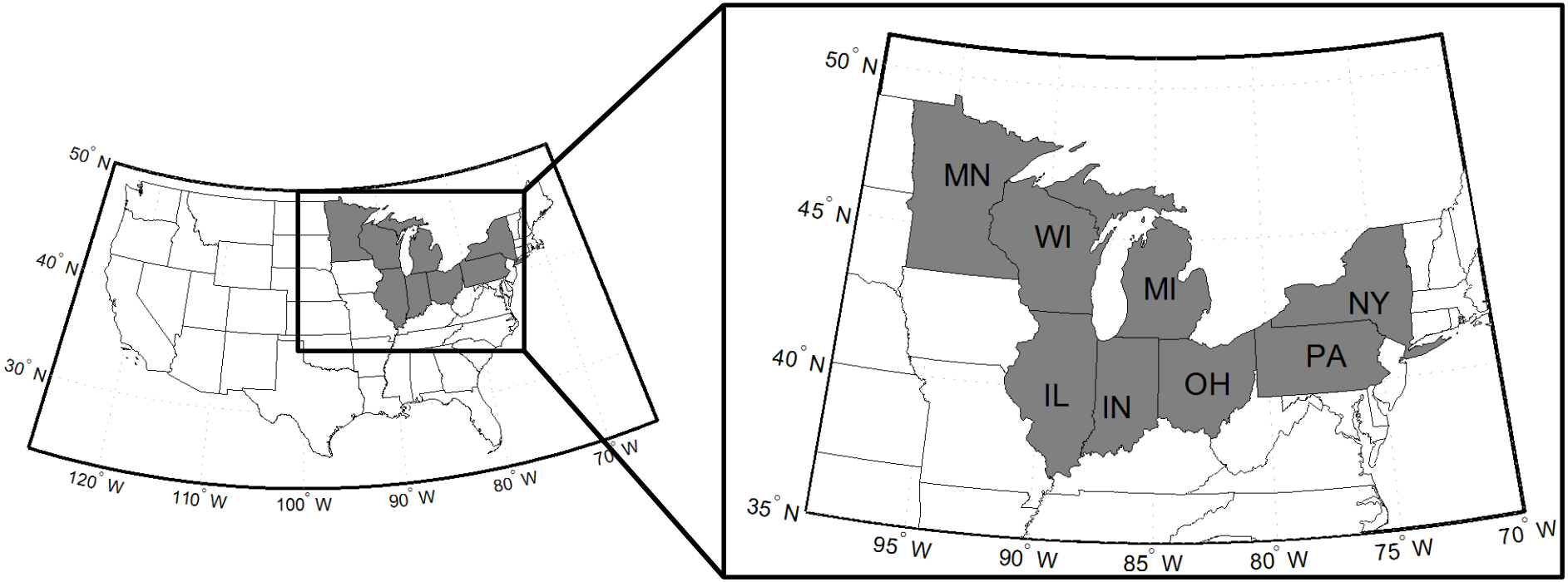
$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_n \|\boldsymbol{\beta}\|_1 \right\}$$

$\|\beta\|_1$: regularization encouraging sparsity, i.e. most elements to be zero



Understanding Factors affecting Precipitation over Great Lakes

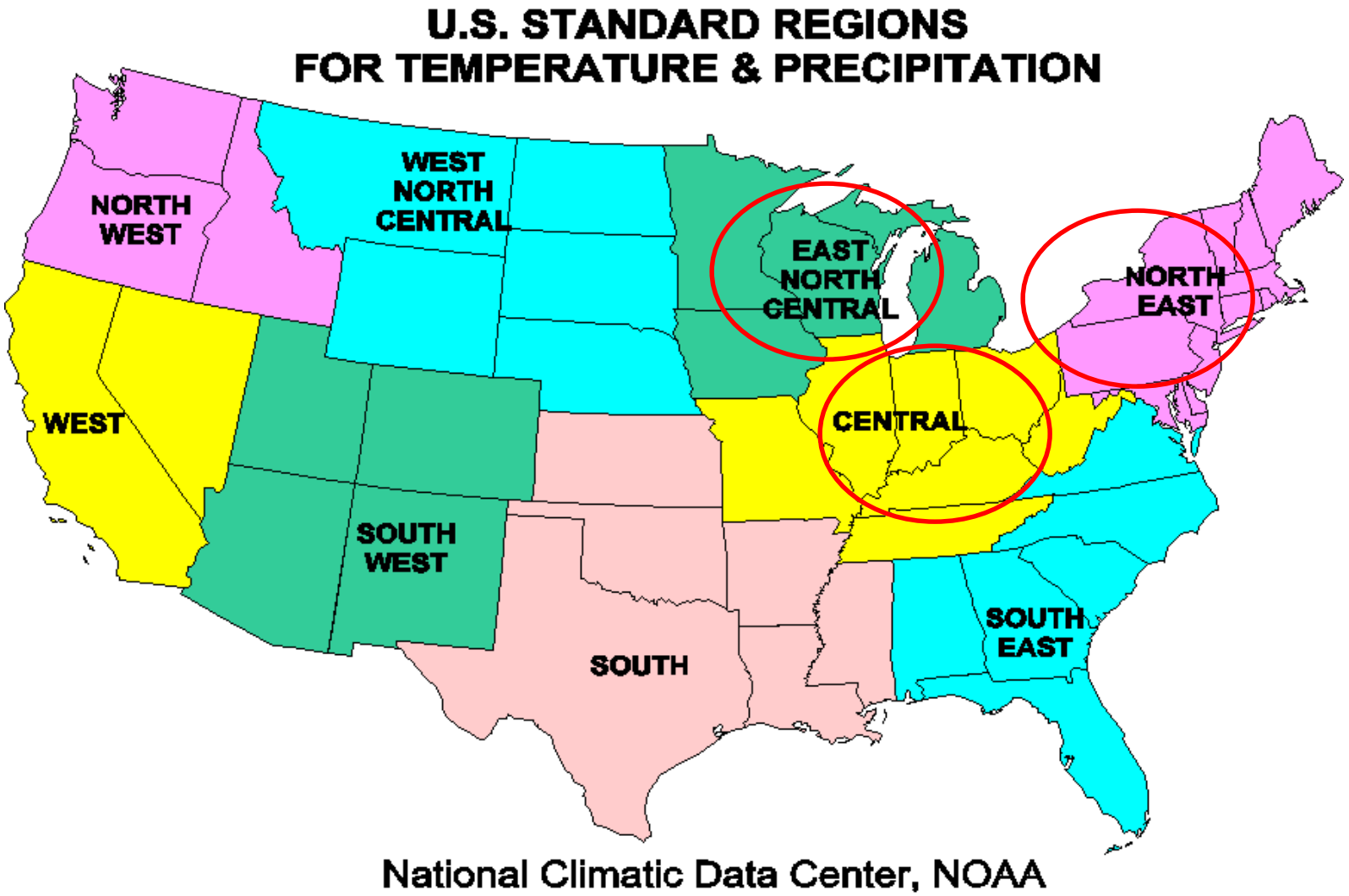
Great Lakes States



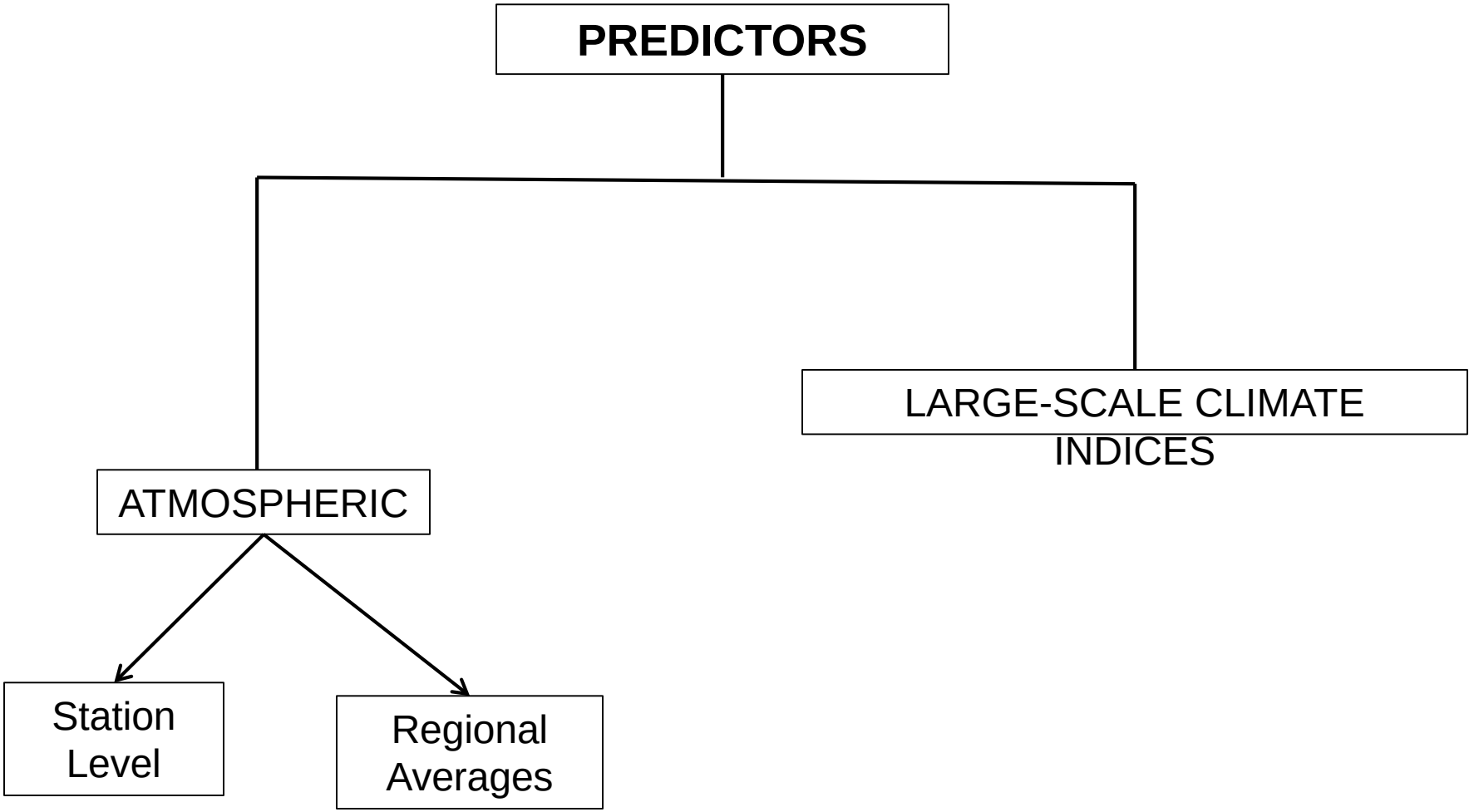
Response: Precipitation

- Station Data (USHCN*) over 8 states (~200 Stations)
  - Near lakes (<200km.)
- Seasonal (DJF) Precipitation
  - Transformed (z-score) to be approximately Gaussian
  - Years: 1979 – 2011

Regions


U.S. STANDARD REGIONS
FOR TEMPERATURE & PRECIPITATION
NORTH WEST
WEST NORTH CENTRAL
EAST NORTH CENTRAL
NORTH EAST
WEST
CENTRAL
SOUTH WEST
SOUTH
SOUTH EAST
National Climatic Data Center, NOAA

Seasonal Covariates



Covariates: Atmospheric Variables

Regional Averages:

- Regional Minimum Temperature (TRegmin),
- Regional Maximum Temperature (TRegmax)
- Regional Mean Air Temperature at 500mb (Reg_AIR_500)

Station Level:

- Winter Minimum Temperature (Tmin),
- Mean Winter Maximum Temperature (Tmax)
- Mean Air Temperature at 500mb (AIR_500)

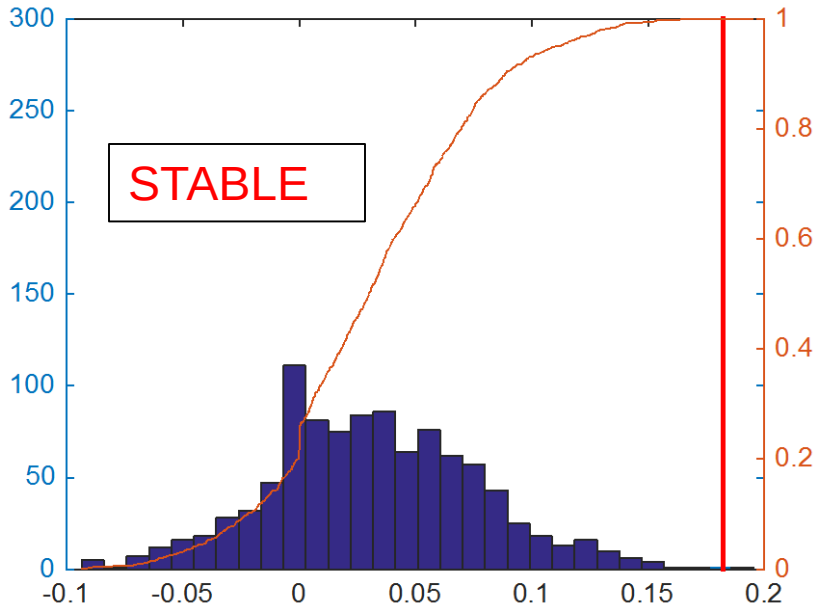
Large-Scale Climate Indices*

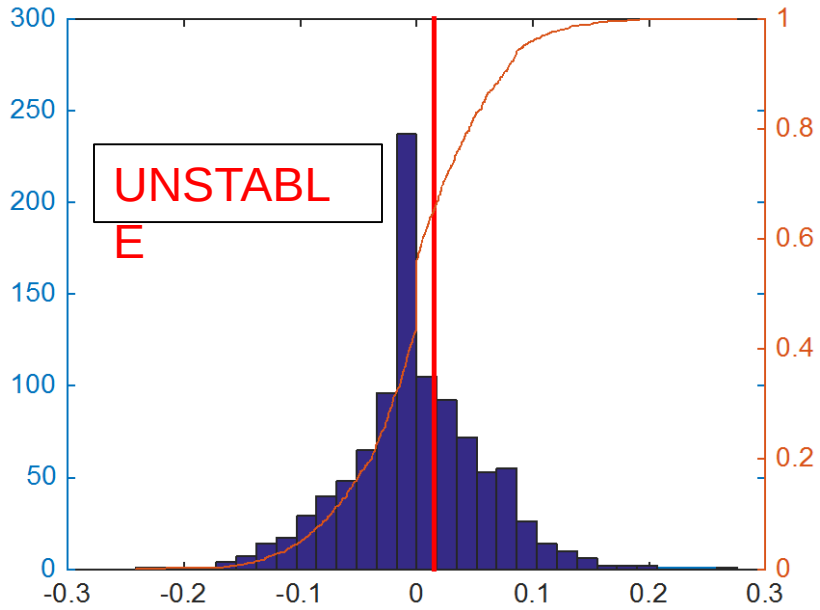
All 12 months as covariates:

1. North Atlantic Oscillation (**NAO**)
2. East Atlantic Pattern (**EA**)
3. West Pacific Pattern (**WP**)
4. East Pacific/North Pacific Pattern (**EPNP**)
5. Pacific/North American Pattern (**PNA**)
6. East Atlantic/West Russia Pattern (**EAWR**)
7. Scandinavia Pattern (**SCA**)
8. Tropical/Northern Hemisphere Pattern (**TNH**)
9. Polar/Eurasia Pattern (**POL**)
10. Pacific Transition Pattern (**PT**)
11. **Nino 1+2**
12. **Nino 3**
13. **Nino 3.4**
14. **Nino 4**
15. Southern Oscillation Index (**SOI**)
16. Pacific Decadal Oscillation (**PDO**)
17. Northern Pacific Oscillation (**NP**)
18. Tropical/Northern Atlantic Index (**TNA**)
19. Tropical/Southern Atlantic Index (**TSA**)
20. Western Hemisphere Warm Pool (**WHWP**)

*CPC Data: http://www.cpc.ncep.noaa.gov/products/MD_index.shtml

LASSO with Random Permutation Test





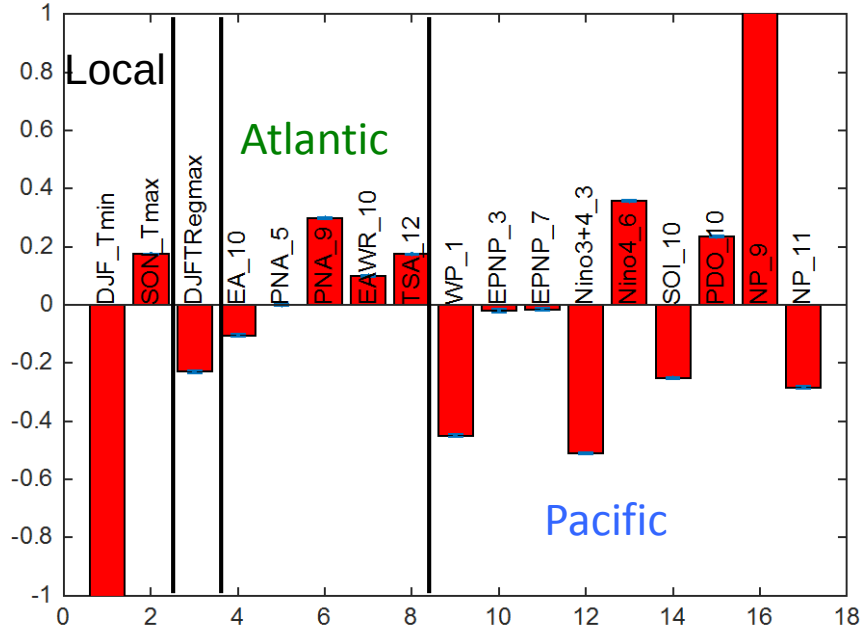
- Dominant feature:
  - non-zero Lasso weight
  - Less than 5% chance of having random large value

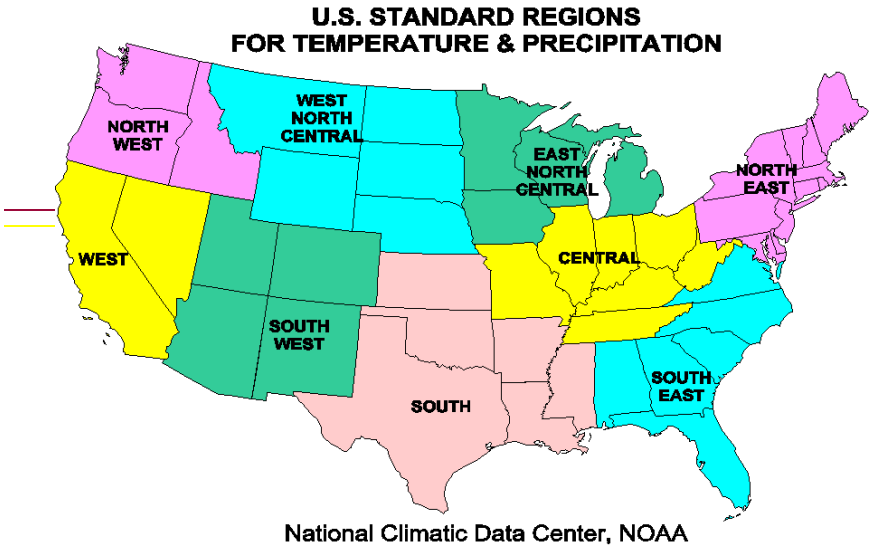
Dominant Features

- Predictive Power?
- Split Data 70/30
- Permutation Test on 70% data
    - Obtain Dominant Features
- Leave-1-out CV on 30 with OLS
    - Climatological Mean
    - ALL 232 Features (Regional+Ocean Indices+Local)
    - Only dominant features

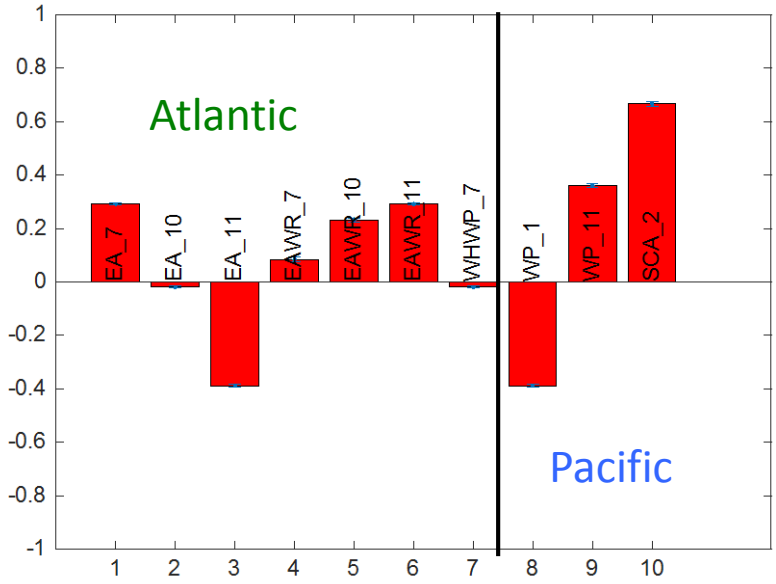
DJF (WINTER) PRECIPITATION

East-North-Central Region

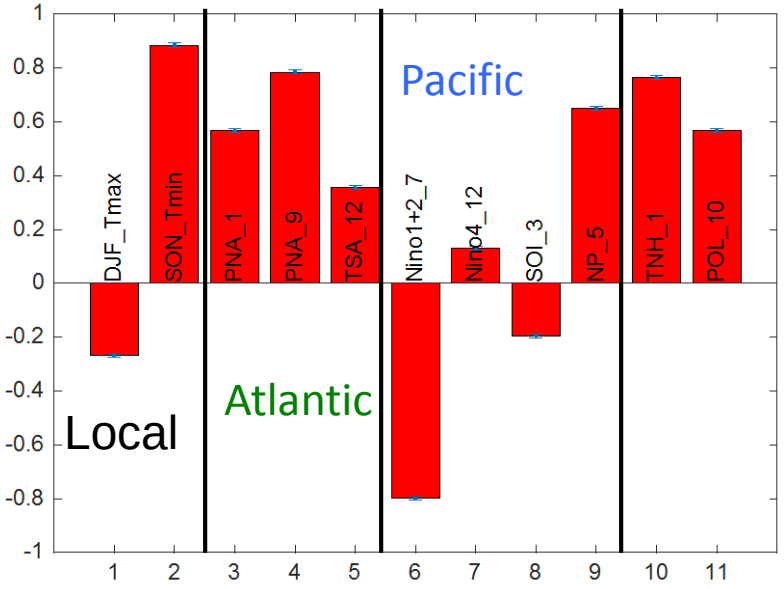




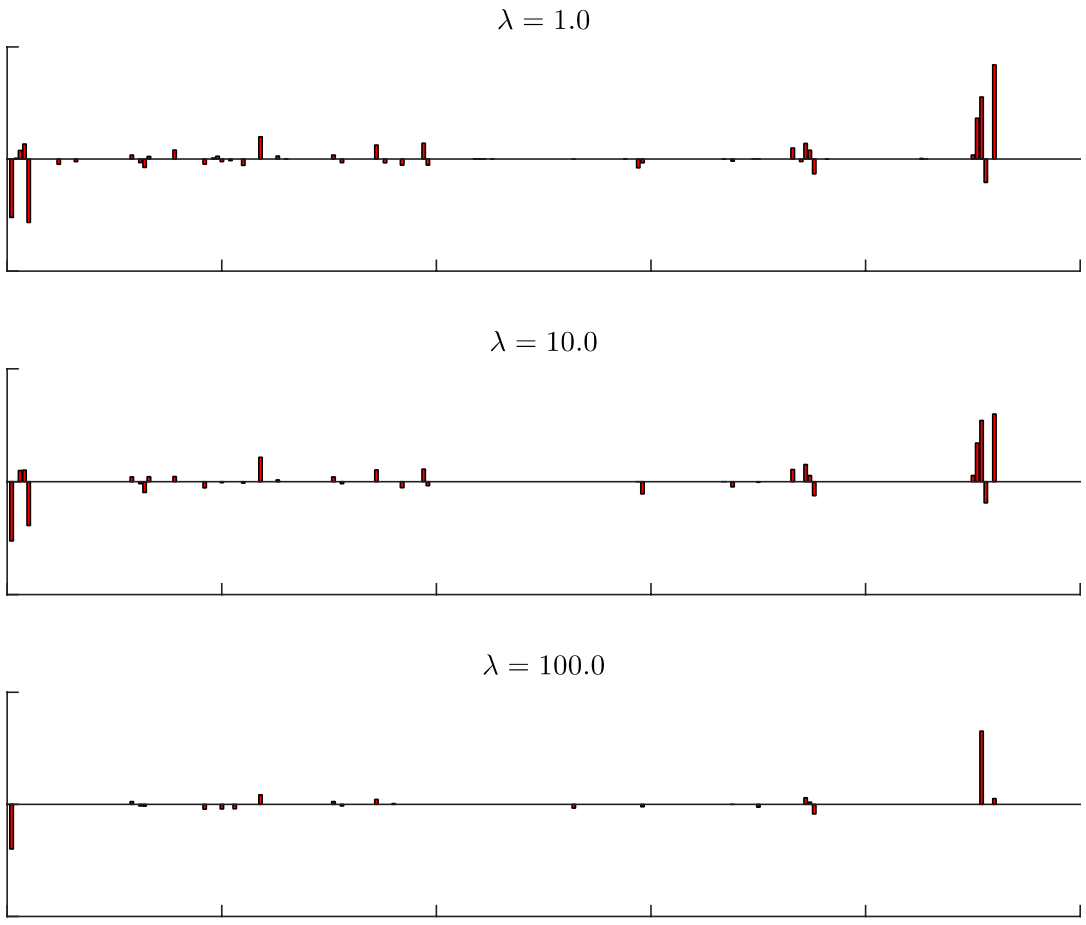
NorthEast Region



Central Region



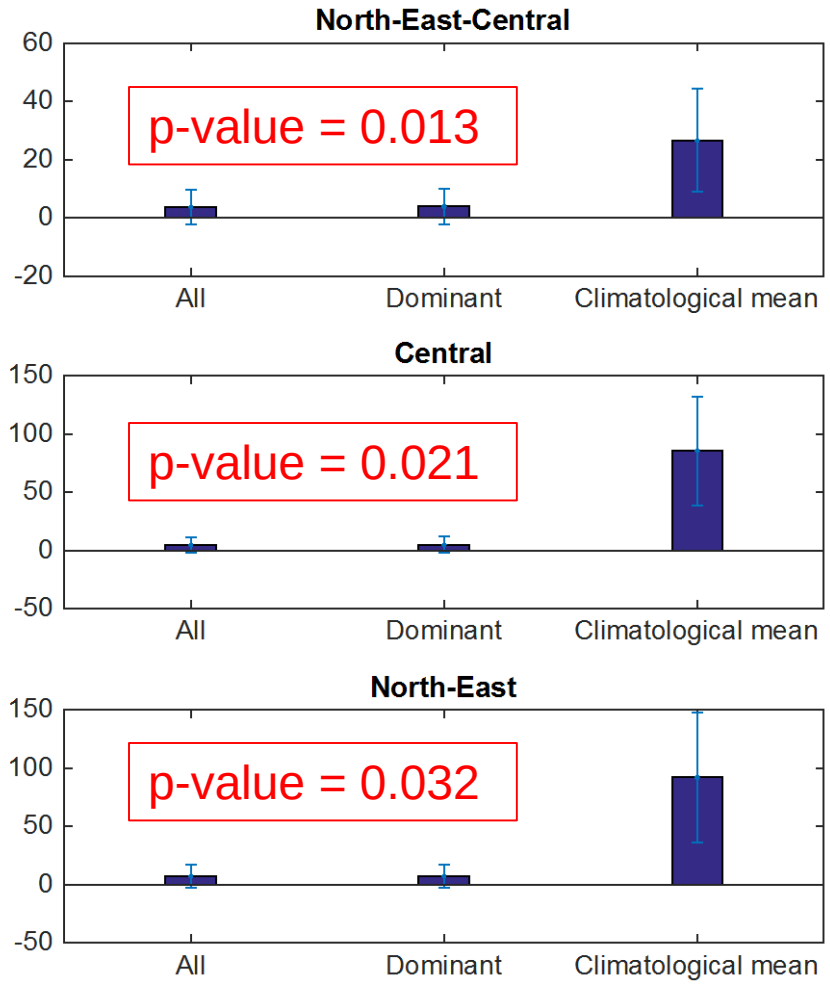
Stable Sets

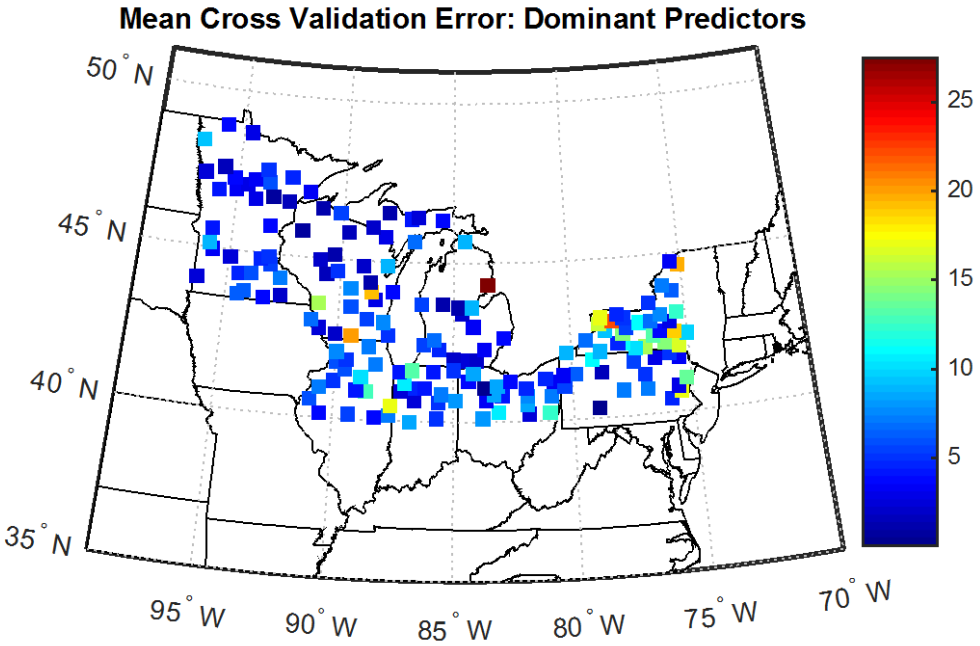


"Pruning" with larger penalty

DJF (Winter) Precip: MSE

MSE: hundredths of an inch

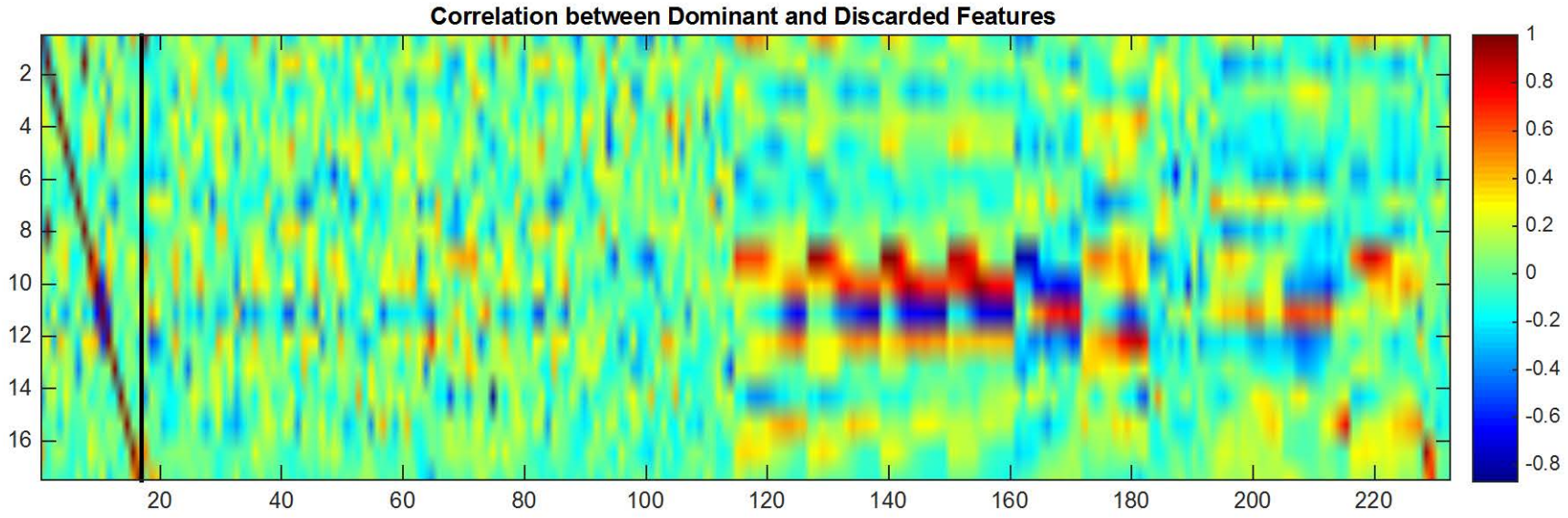




All = ALL 232 features
Dominant = Dominant features ONLY

Data Subspace

Cross-Correlation between Dominant and Discarded Features



Dominant features

Discarded features

- # of Dominant Features = 17

Summary

- “Stable” Covariates ⟷ Dominant Features
- Hypothesis Testing
  - Permutation Test
- Predictive power
- Future Directions
  1. Non-Gaussianity: GLM’s
  2. Spatio-Temporal Dependencies
  3. Understanding Processes: Representation

Acknowledgements

Collaborators

- Stefan Liess (UMN)
- Arindam Banerjee (UMN)
- Debasish Das (Verizon)
- Auroop Ganguly (NEU)

Funding

Thanks