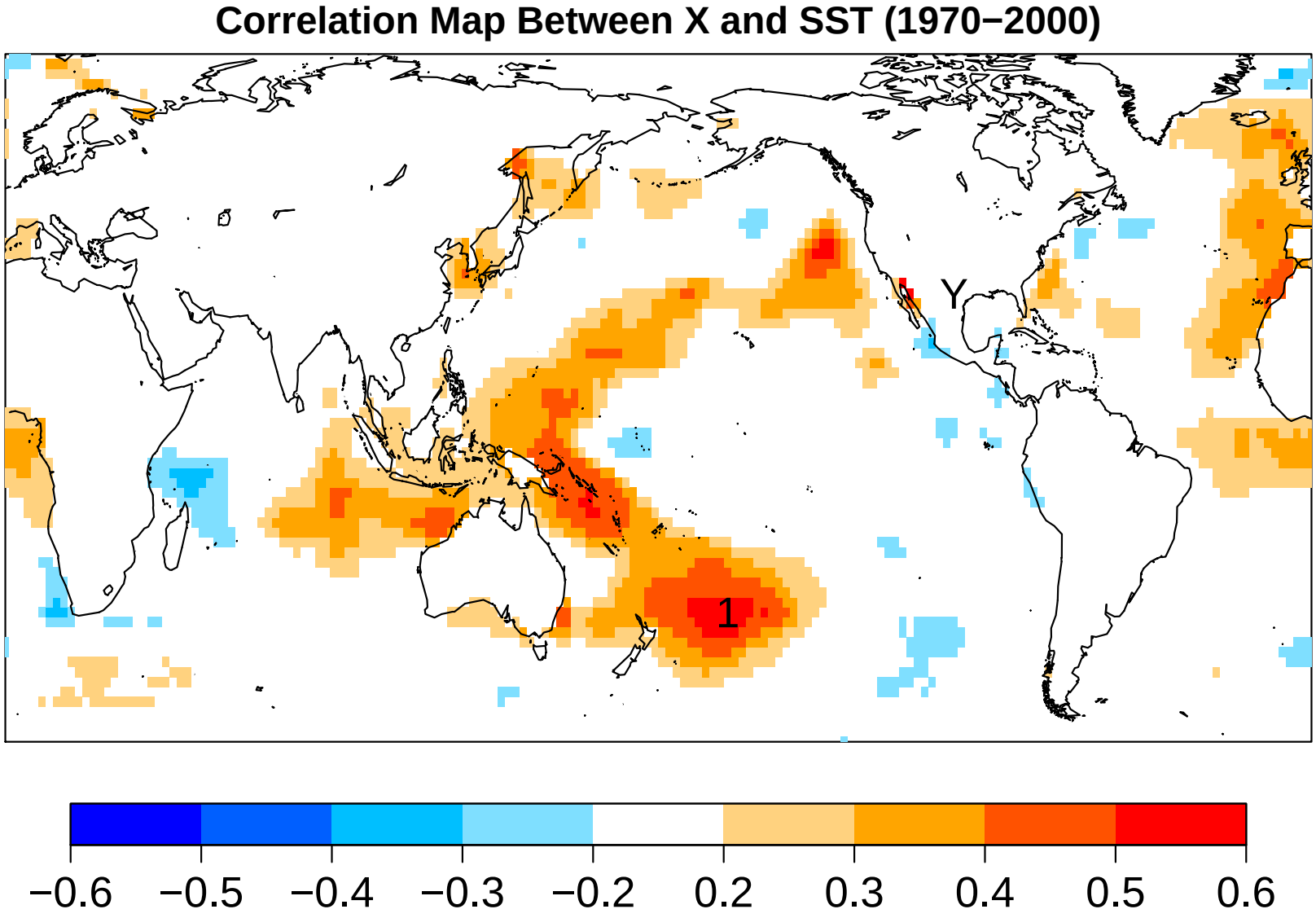
Finding Relations in Climate Data Sets

Timothy DelSole

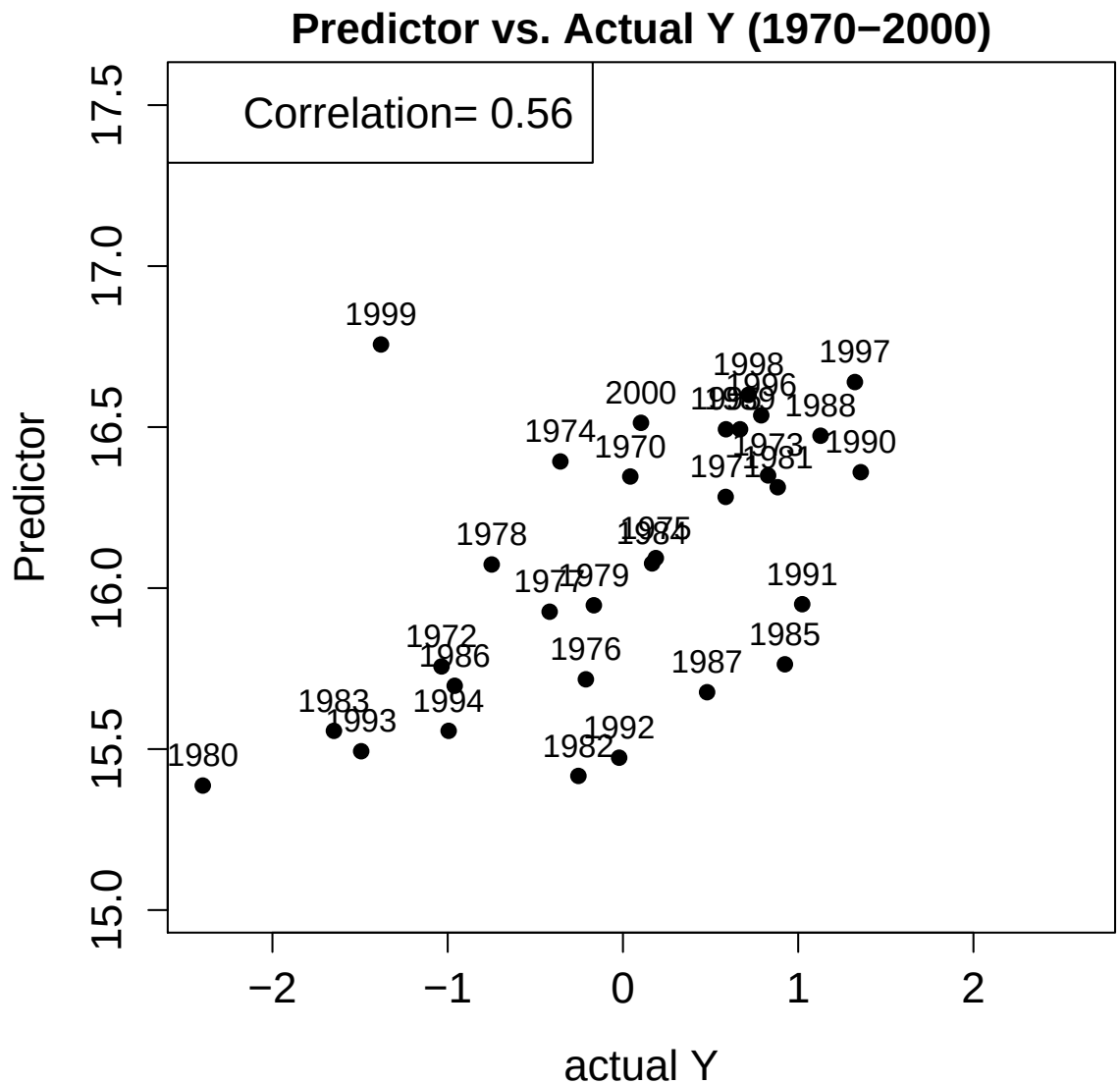
George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

August 5, 2015

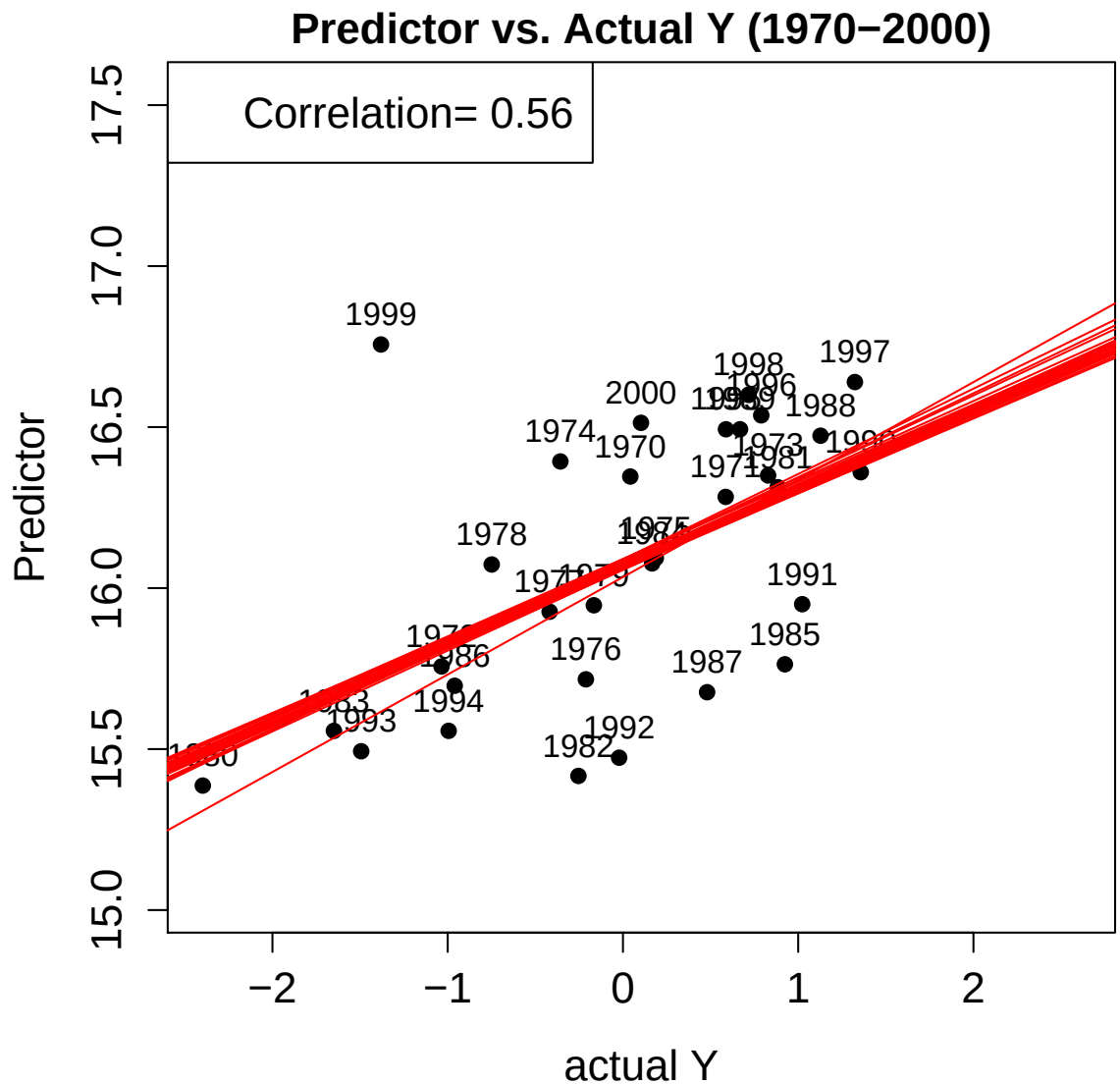
collaborators: Michael Tippett, Arindam Bannerjee, Claire Monteleoni


Correlation Map Between X and SST (1970−2000)
Y
1
−0.6
−0.5
−0.4
−0.3
−0.2
0.2
0.3
0.4
0.5
0.6

Scatter Plot



Leave-One-Out Regression Fits



Surprise! Y is random!

Surprise! Y is random!

- On a $2.5^\circ \times 2.5^\circ$ grid, ocean surface is about 8000 points.
- Only about 50 years of observations of global ocean surface.
- Physics varies with season, so seasons cannot be pooled.

Estimate 8000 parameters using 50-150 data points

This is an example of data fishing.

1. Not all climate scientists understand this issue.
2. For those who do, machine learning "looks" like data fishing.
3. It is useful to challenge climate scientists about how they know what they tell you.
4. Some correlations are trusted.

ENSO Teleconnections?

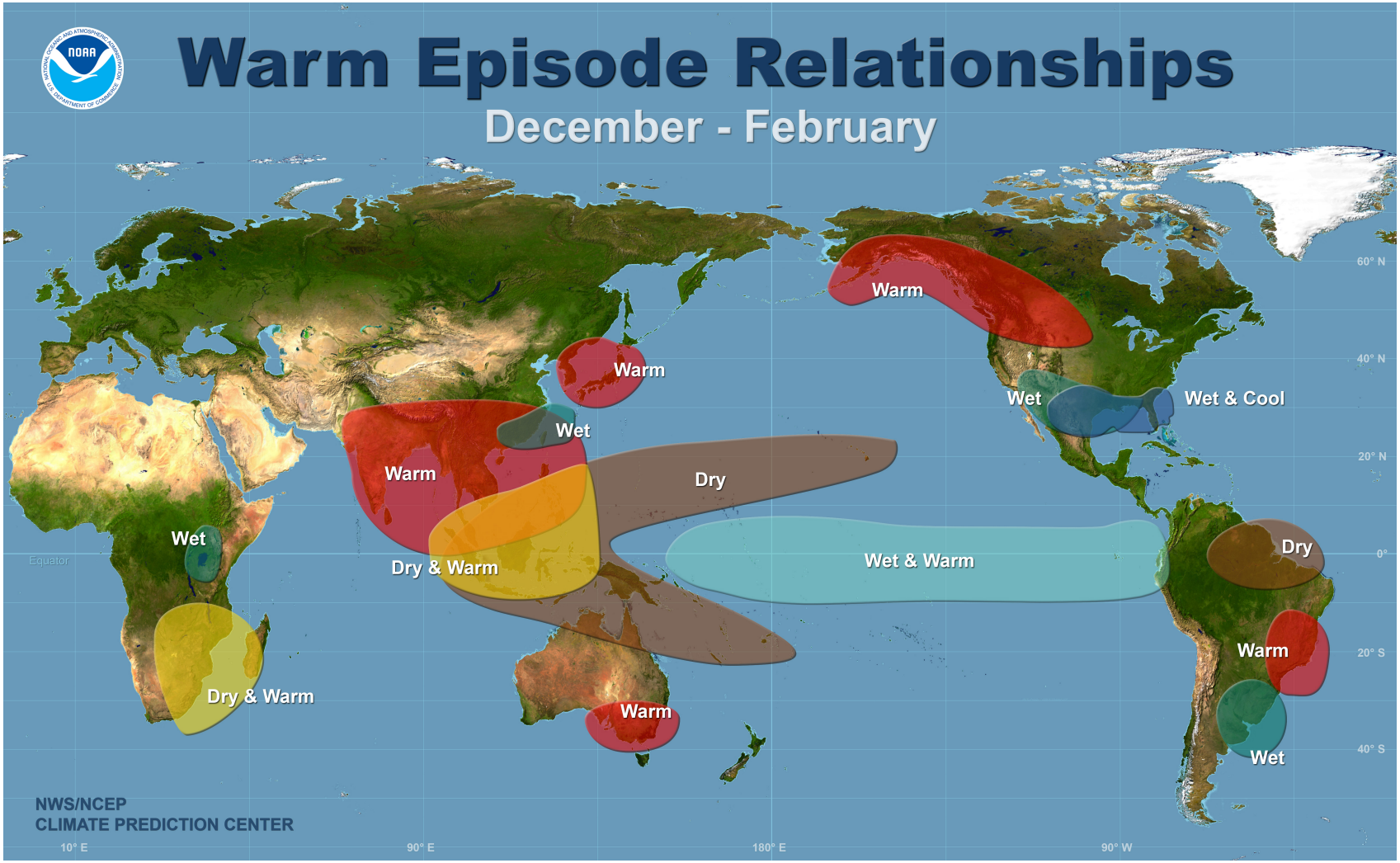


figure http://www.cpc.noaa.gov/products/analysis_monitoring/impacts/warm.gif

Proving an Observed Relation is Real

- Relation holds in independent data.
- Relation can be reproduced by climate models.
- Relation can be understood through simple dynamical models.

ENSO Teleconnection Inferred in 1989 (Precipitation)

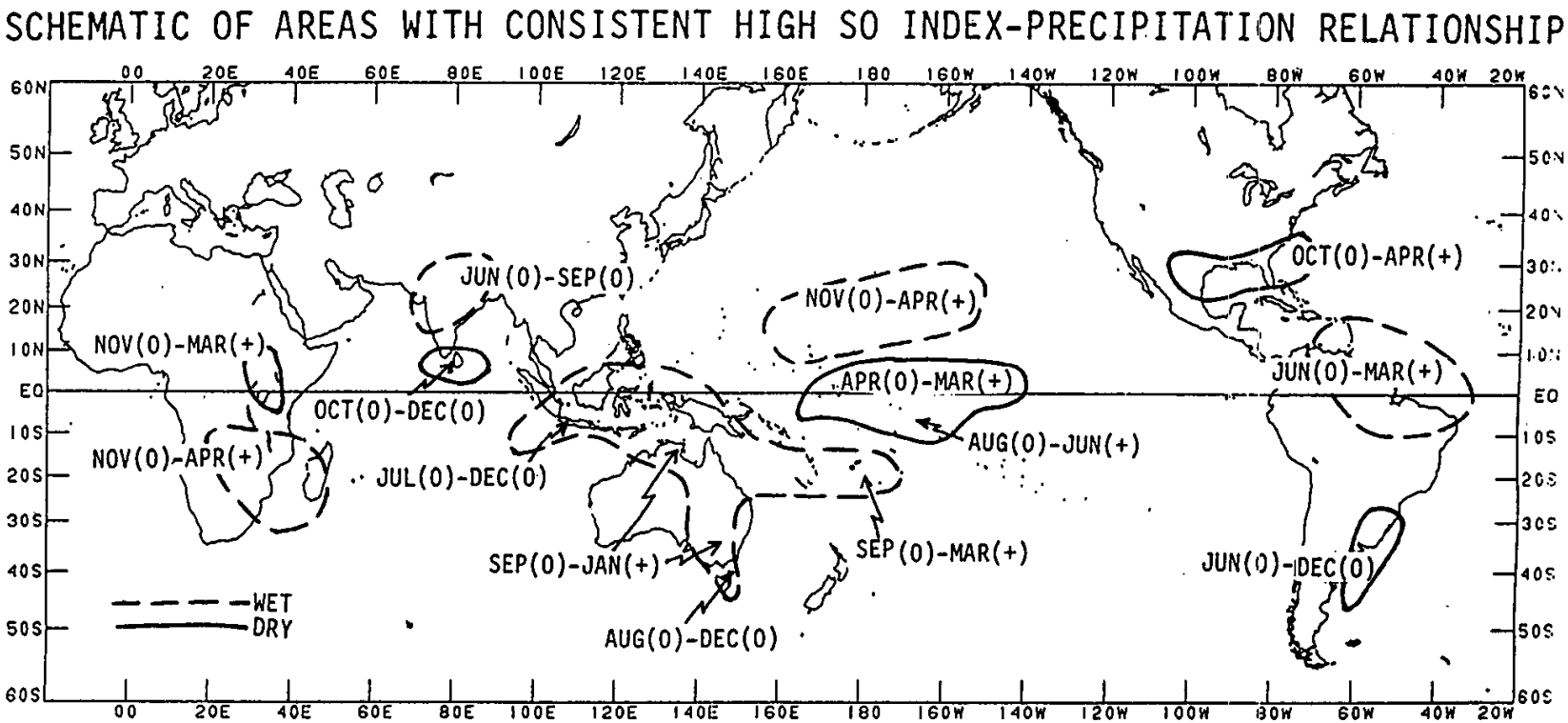


FIG. 18. Schematic representation of the principal areas of high-SO index related precipitation based on the analysis of precipitation composites and time series. Regional maps should be consulted for details.

figure: Ropelewski and Halpert (1989, J. Climate)

Machine Learning Approach: Bet on Sparsity

Frame the estimation such that most parameters are zero.

"Use a procedure that does well in sparse problems, since no procedure does well in dense problems."

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*, 2nd edition

---

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*, 2nd edition

Examples of Sparsity

grid points:	variable is related to localized geographic regions.
matrix rank:	variable is related to a few patterns.
spectral:	variable fluctuates over a small range of frequencies
PCs:	variable is related to components with high variance
variables:	variable is related to certain physical variables

What is the most appropriate form of sparsity when finding relations in climate data?

Small-scale structure is less trustworthy than large-scale structure in interseasonal climate.

Climate Models

- Accuracy of numerical solutions of partial differential equations degrades with decreasing length scale.
- Subgrid scale phenomena are not solved directly, but are parameterized in terms of large scale grid quantities.
- Surface topography is rarely accurately included, and often induces numerical phenomena that have no counterpart in nature.

Eigenvectors of Laplace Operator provide a natural basis set for representing large spatial scales or long temporal scales.

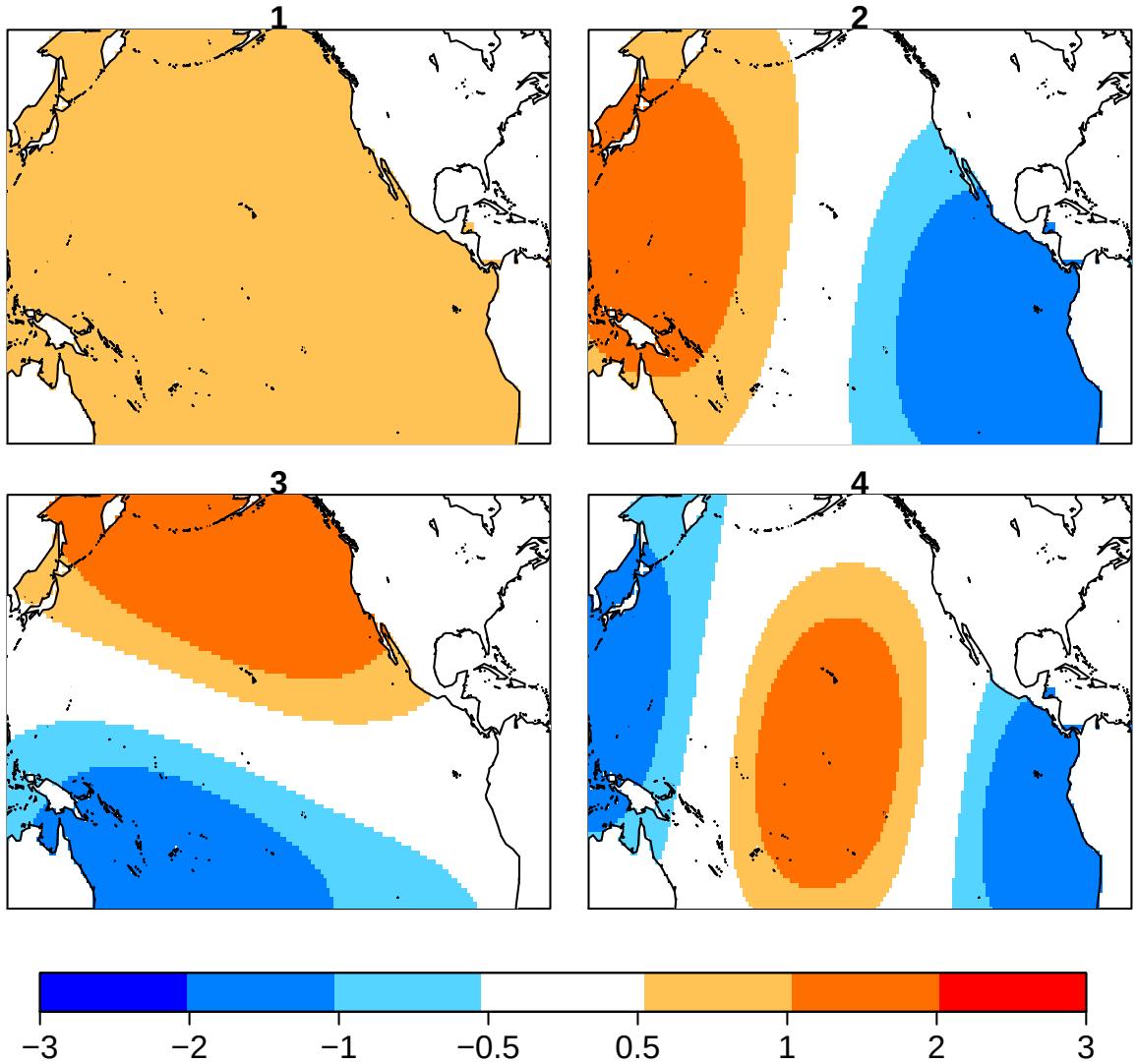
$$\nabla^2 \psi = -\lambda^2 \psi$$

Circle: Fourier series (λ is the frequency)

Sphere: Spherical Harmonics (λ is the total wavenumber)

Large-scale $\rightarrow$ most amplitudes vanish.

Laplacian Eigenvectors Over the Pacific



DelSole and Tippett (2015, *Journal of Climate)*

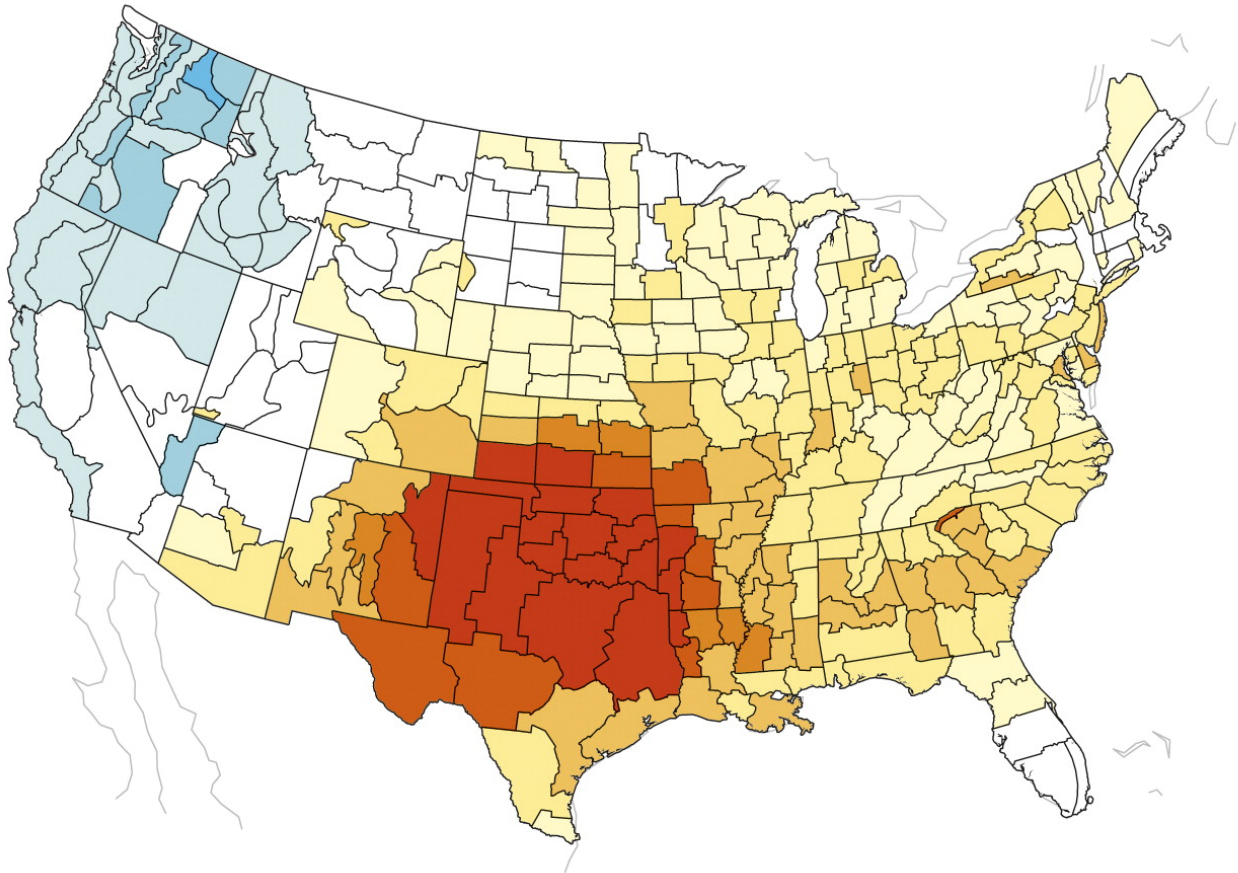
Regularized Linear Regression

$$\| \underbrace{\mathbf{y}}_{\textit{predictand}} - \underbrace{\mathbf{X}}_{\textit{predictors}} \underbrace{\mathbf{w}}_{\textit{weights}} \|^2 + \underbrace{\lambda R(\mathbf{w})}_{\textit{regularizer}}$$

Typical regularizers (choice of R) in machine learning:

- $w_k = 0$ for $k \geq K$ (Principal Components Regression)
- L_1 norm ($|\mathbf{w}|$) (LASSO)
- L_2 norm ($\|\mathbf{w}\|^2$) (Ridge)

Observed 2011 JJA Temperature



Degrees Celsius (°C)

−3 −2.5 −2 −1.5 −1 −0.5 0.5 1 1.5 2 2.5 3

from figure 1 of Hoerling et al. 2013, J. Climate

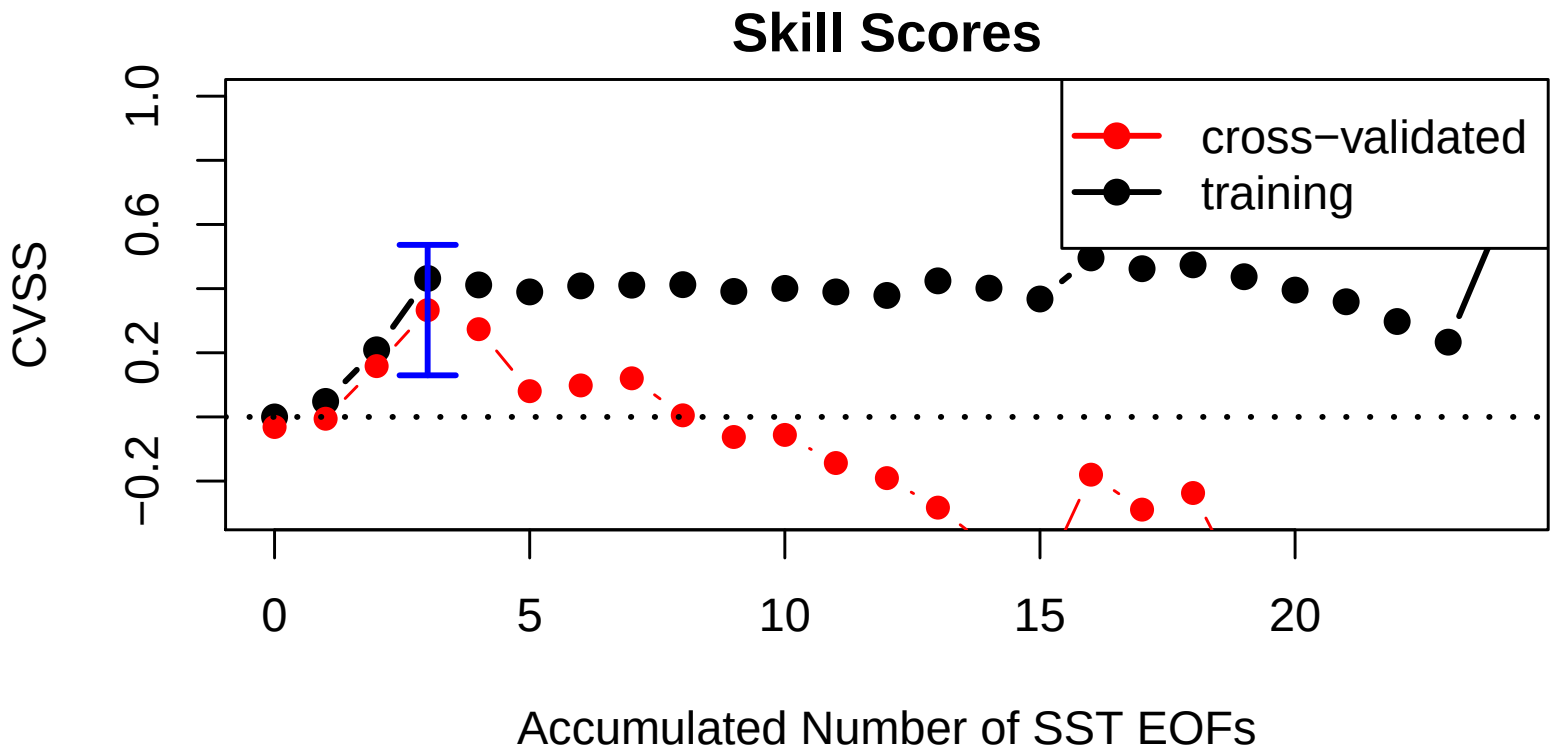
Regression model:

$$\text{Texas temperature} = \text{SST in Pacific} * \text{weights} + \text{noise}$$

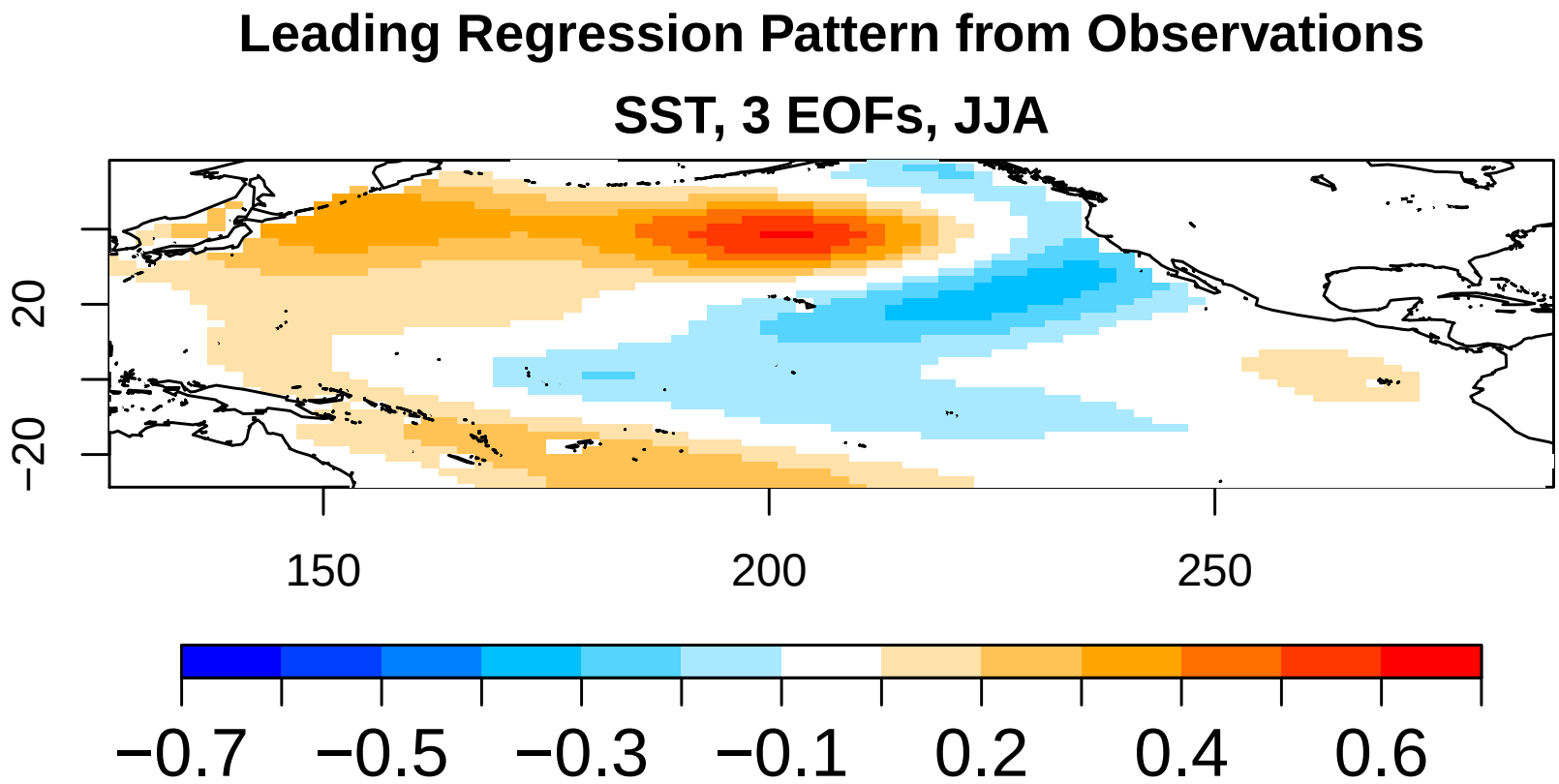
Prediction Measure:

$$\text{Cross-Validated Skill Score} = 1 - \frac{MSE}{\text{variance of Texas Temperature}}$$

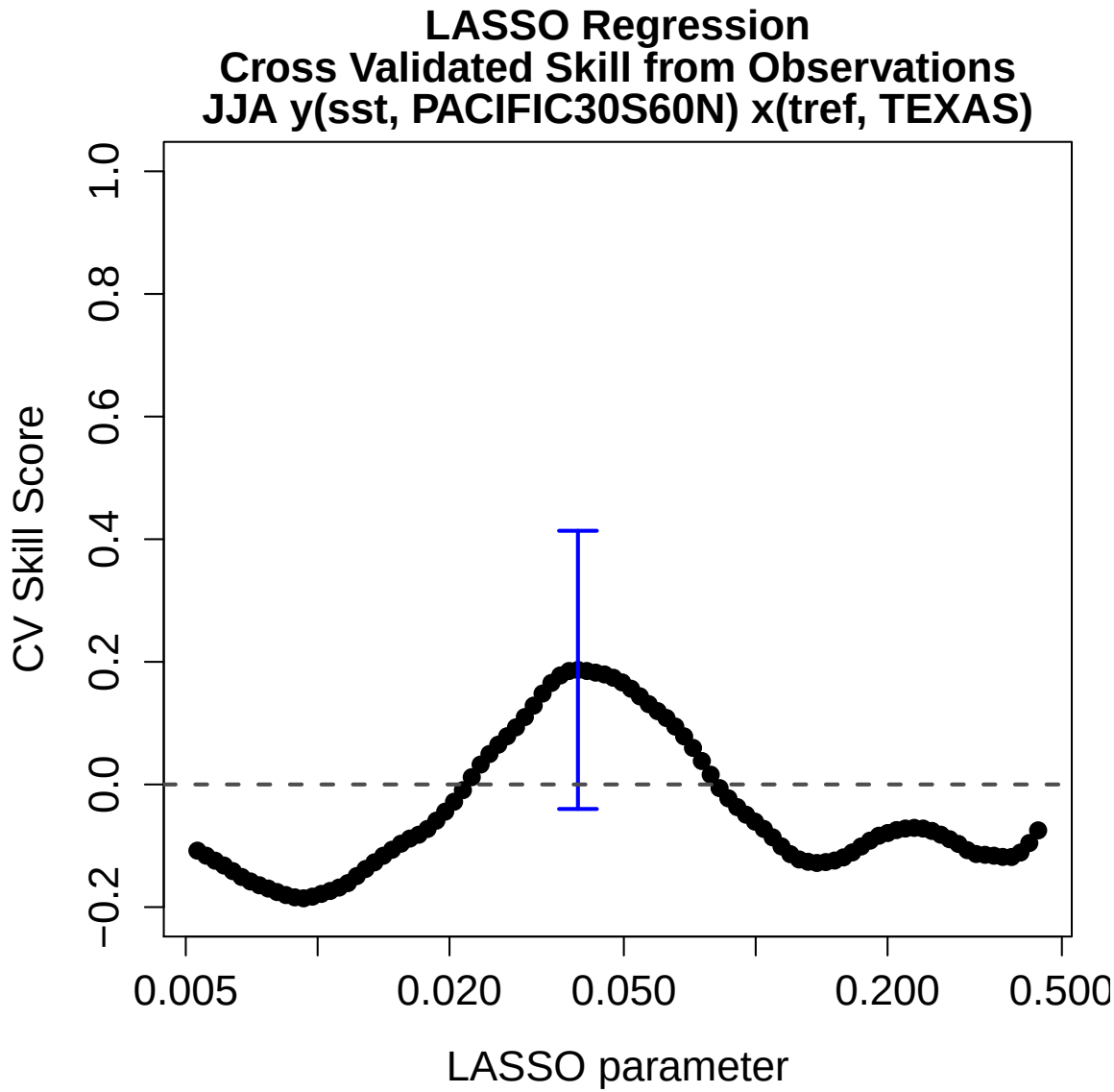
Principal Components Regression



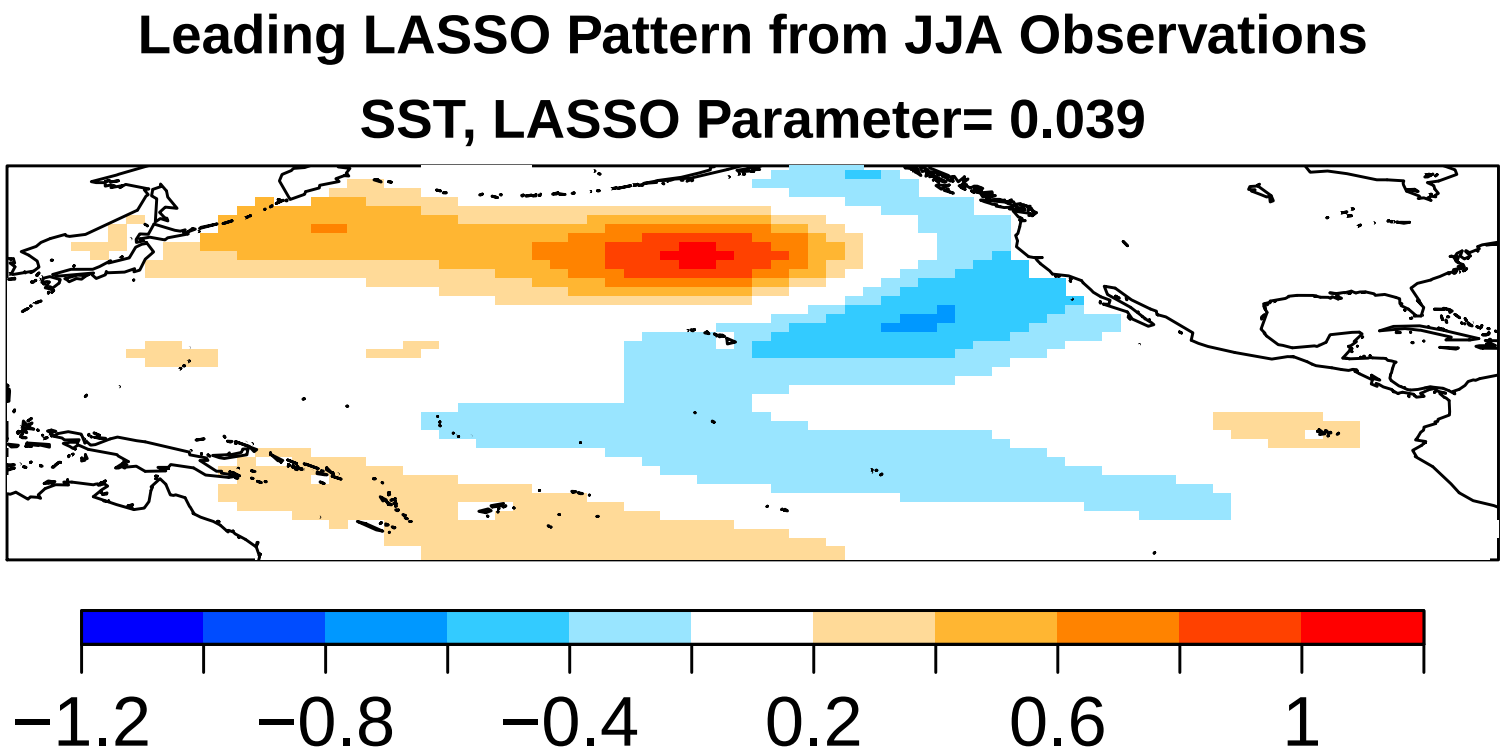
Principal Components Regression



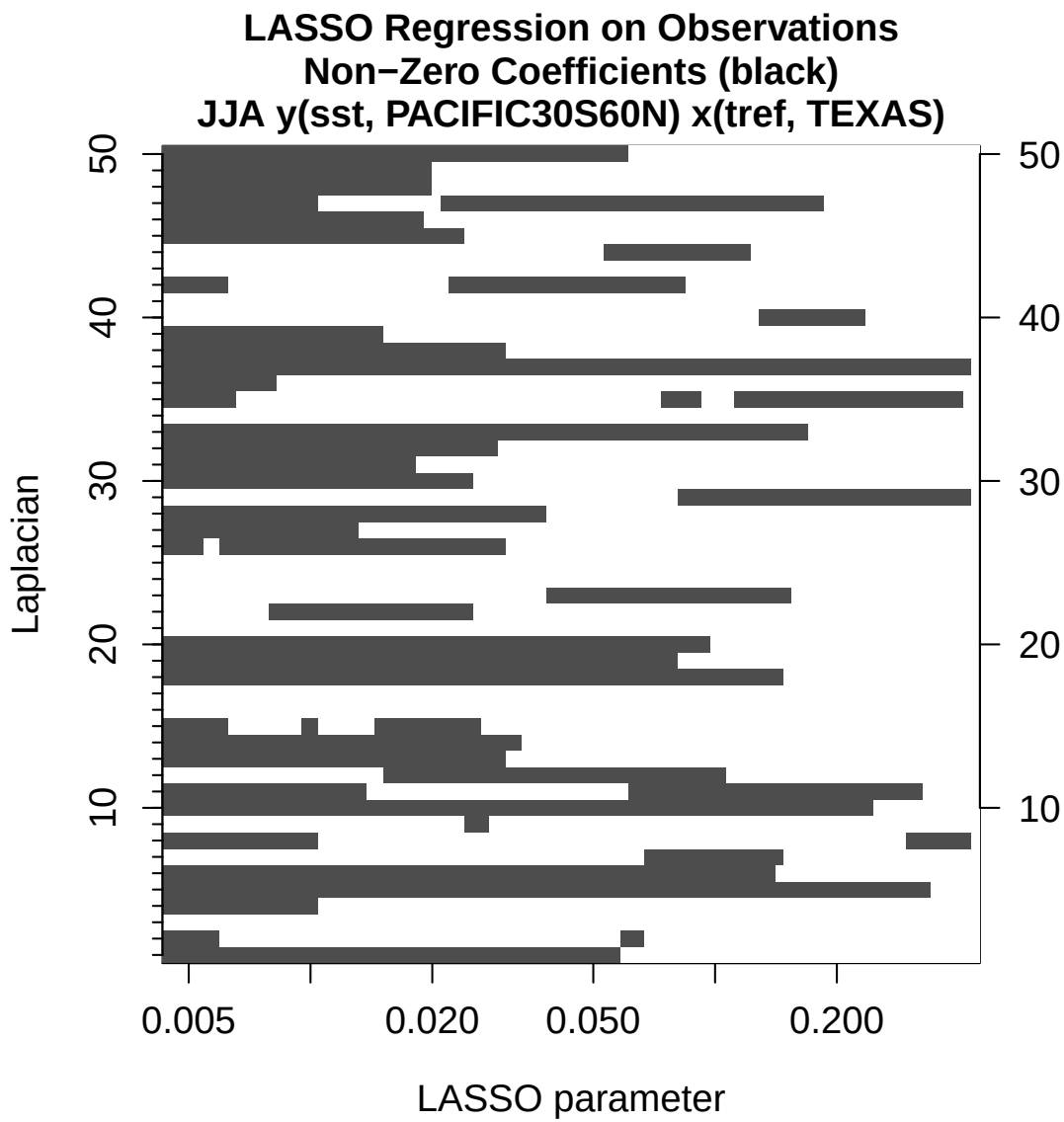
LASSO w/ Laplacians



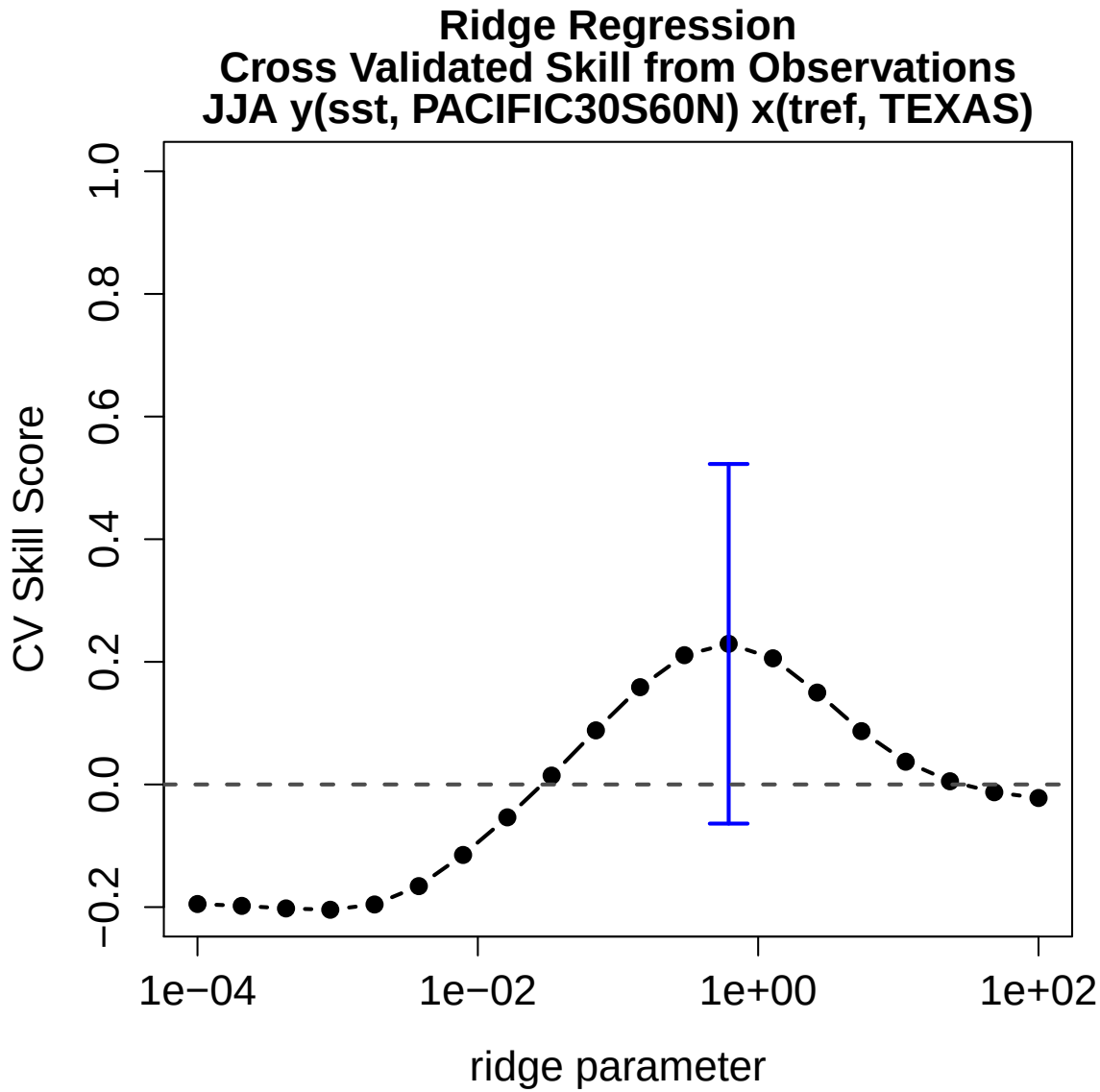
LASSO w/ Laplacians



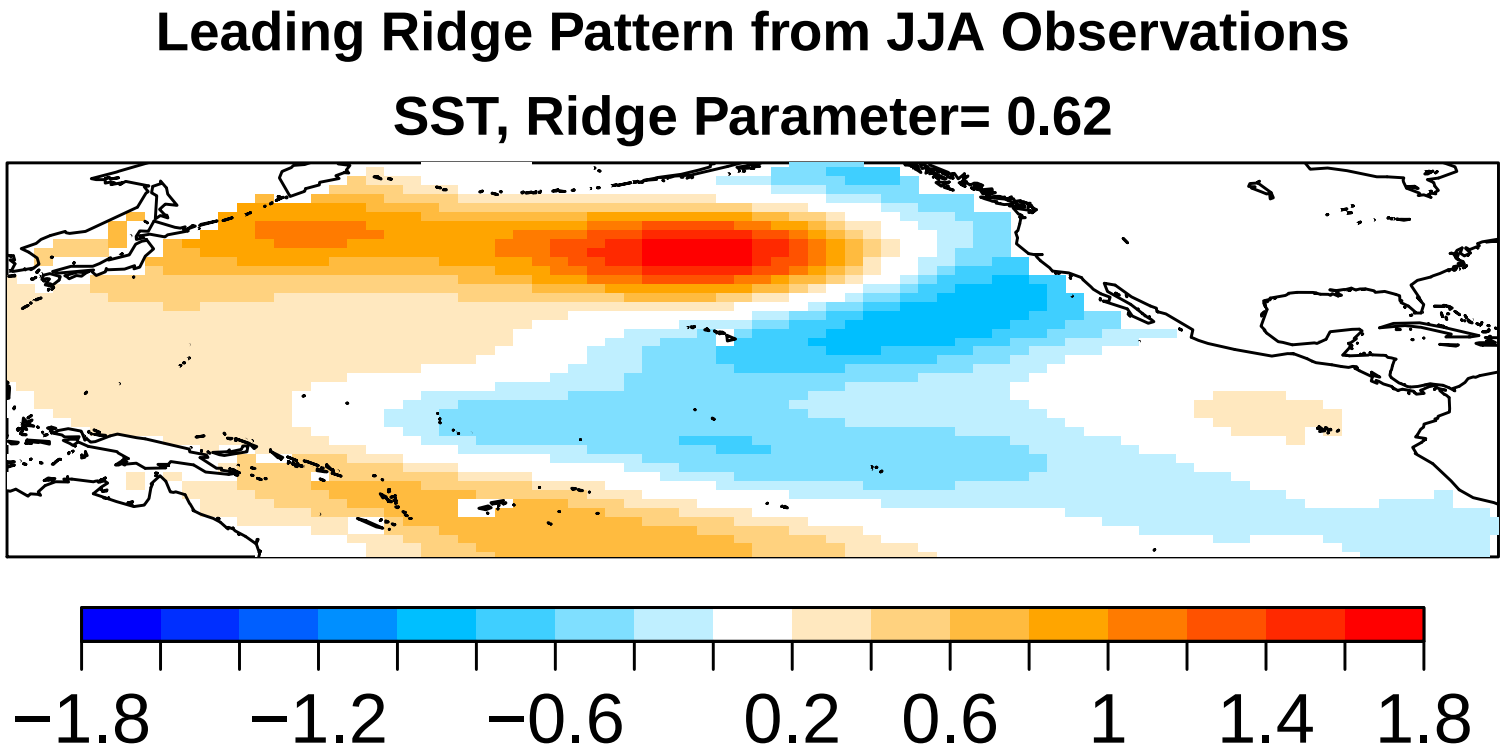
LASSO w/ Laplacians



Ridge w/ Laplacians



Ridge w/ Laplacians



Can Climate Models Reproduce These Relations?

North American Multi-Model Ensemble (NMME)

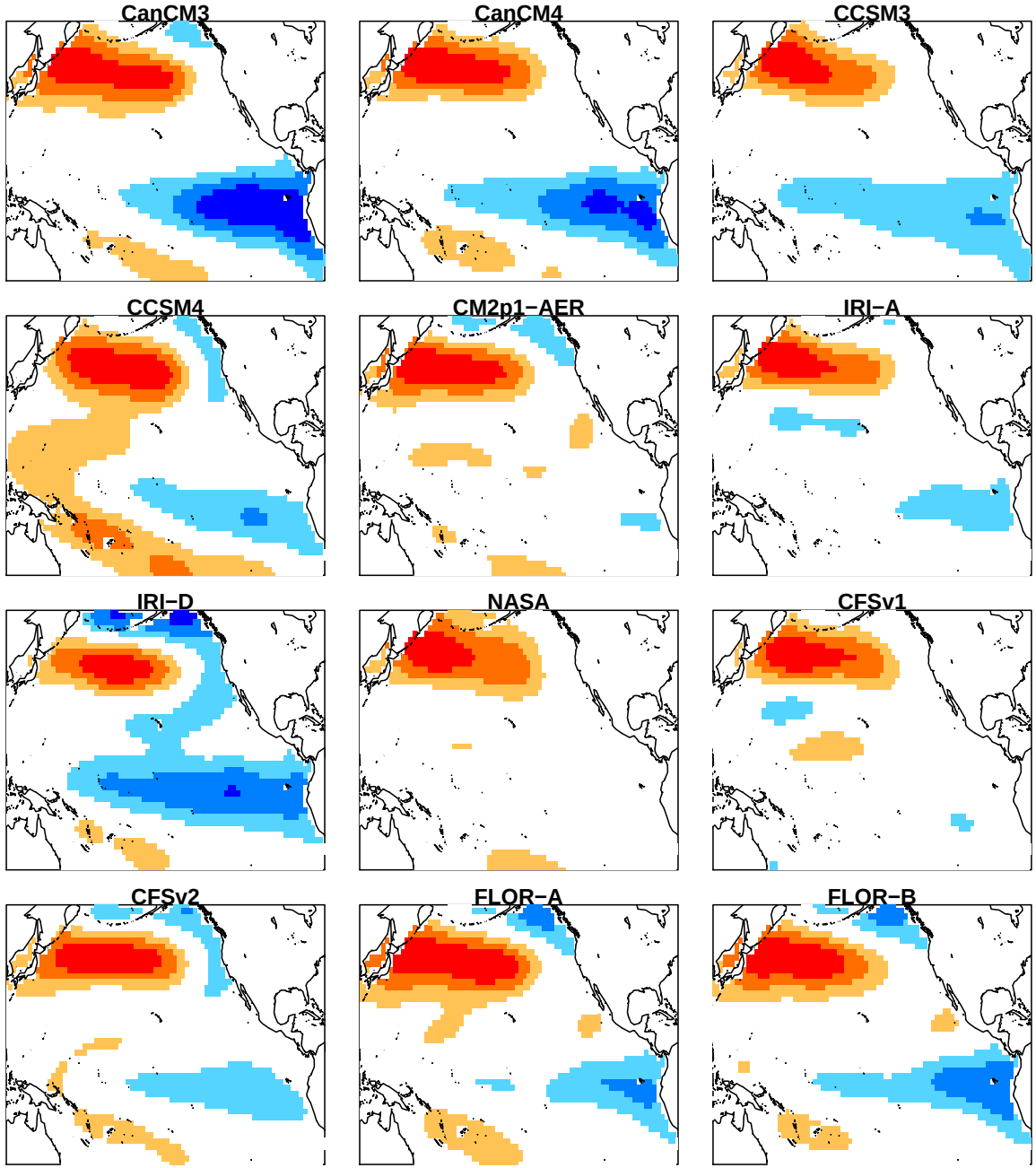
- Seasonal-to-intraseasonal predictions by state-of-the-art models

CFSv1	NOAA Climate Forecast System version 1
CFSv2	NOAA Climate Forecast System version 2
CM2p1-AER	GFDL Climate Model version 2.2
FLOR-A	GFDL
FLOR-B	GFDL
NASA	NASA Goddard Observing System v5
IRI-D	IRI-ECHAM4 Direct Coupling
IRI-A	IRI-ECHAM4 Anomaly Coupling
CCSM4	NCAR Community Climate System Model
CCSM3	NCAR Community Climate System Model
CMC1	Canadian Coupled Climate Model
CMC2	Canadian Coupled Climate Model

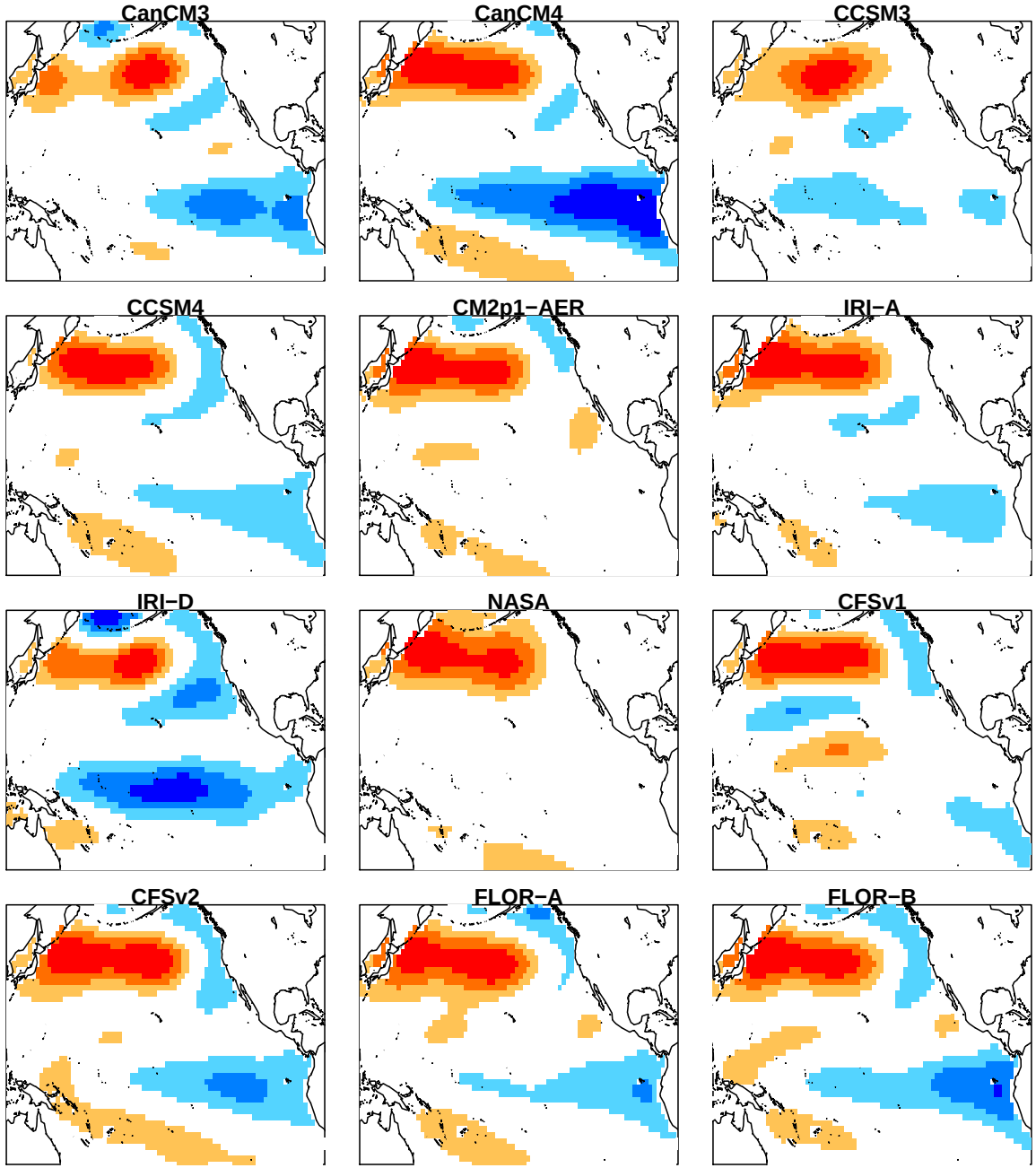
- at least 6 ensemble members per model
- 8-12 month predictions/hindcasts
- 1982-2014 (not all models available during this period)
- pool initial start months January - May

Kirtman et al, 2014, *Bulletin of the American Meteorological Society*

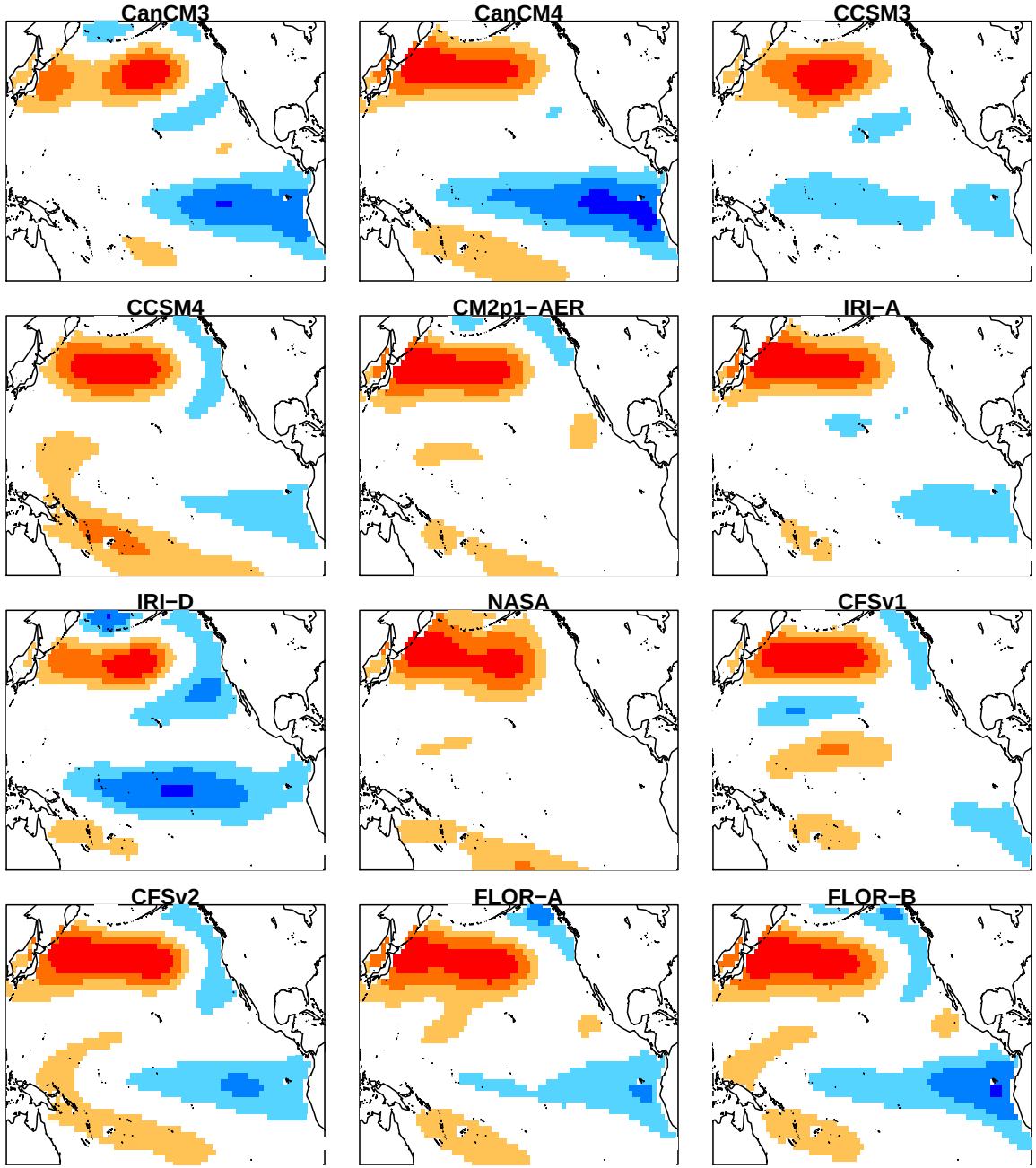
Principal Components Regression



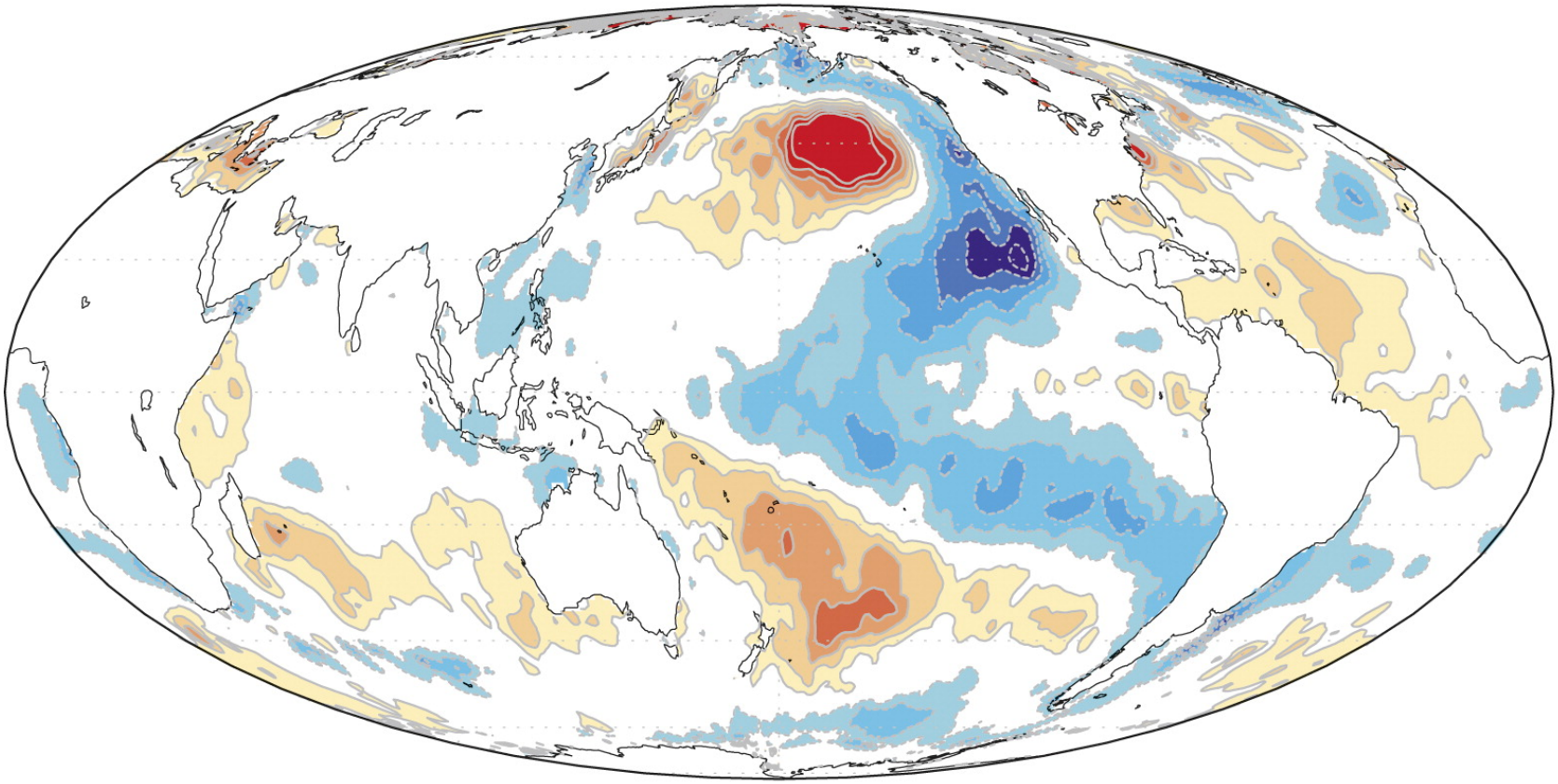
LASSO


CanCM3
CanCM4
CCSM3
CCSM4
CM2p1−AER
IRI−A
IRI−D
NASA
CFSv1
CFSv2
FLOR−A
FLOR−B

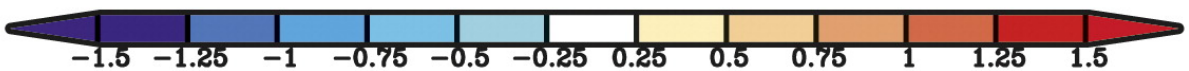
Ridge


CanCM3
CanCM4
CCSM3
CCSM4
CM2p1−AER
IRI−A
IRI−D
NASA
CFSv1
CFSv2
FLOR−A
FLOR−B

Observed SST JJA 2011



Degrees Celsius

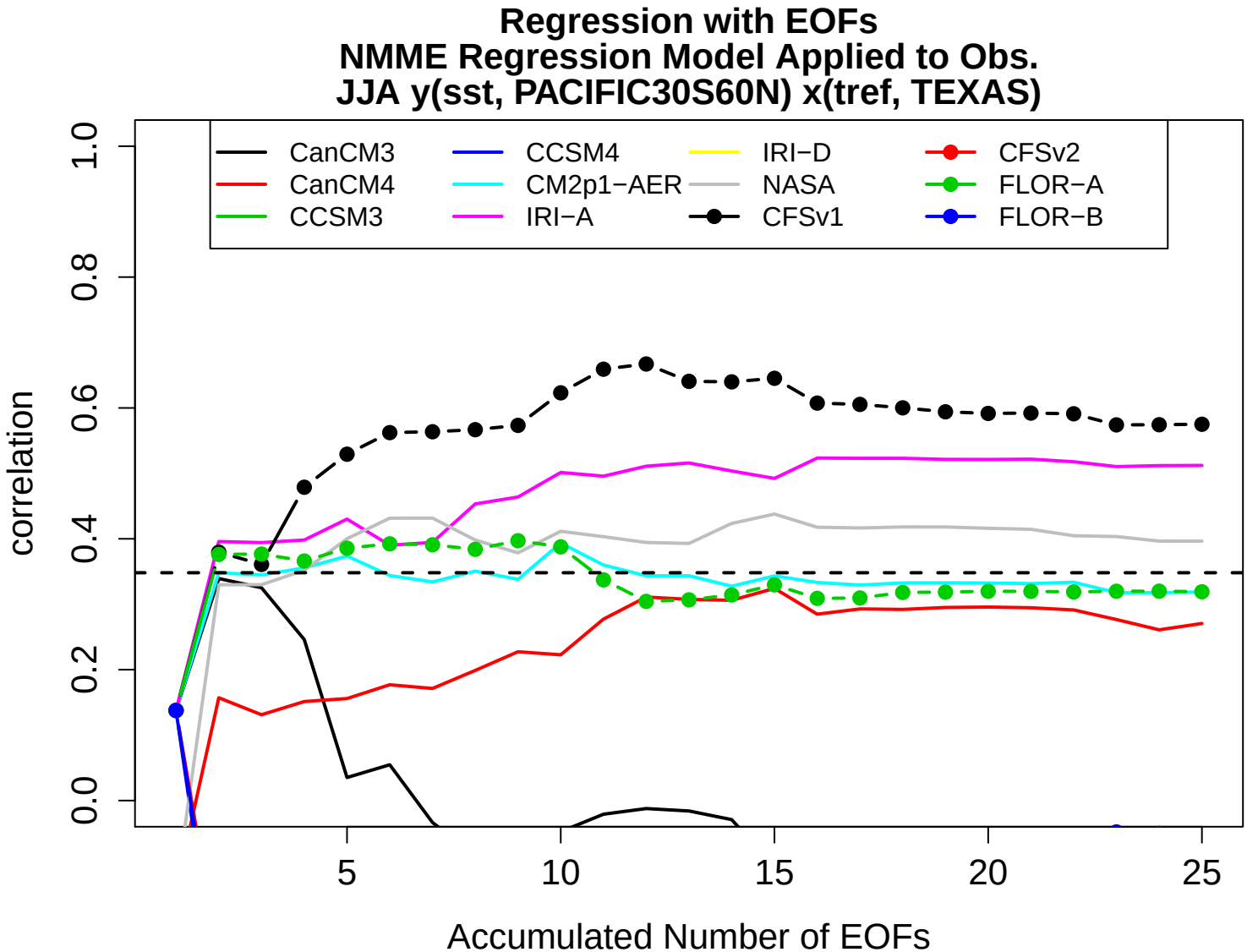


from figure 3 of Hoerling et al. 2013, J. Climate

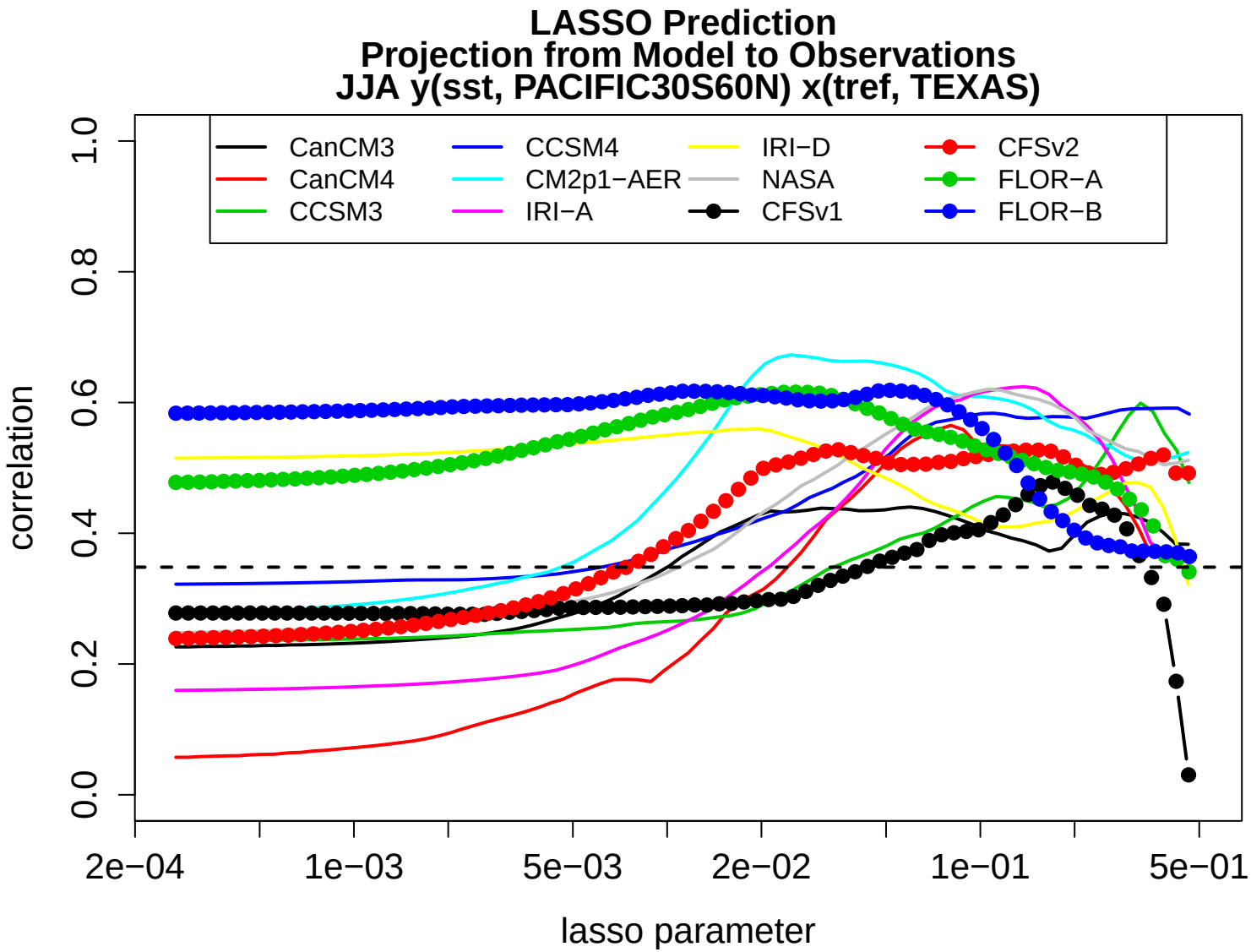
Use relations derived from climate models on observations

$$\underset{\text{observed}}{\text{Texas temperature}} = \underset{\text{observed}}{\text{SST in Pacific}} * \underset{\text{model}}{\text{weights}} + \text{noise}$$

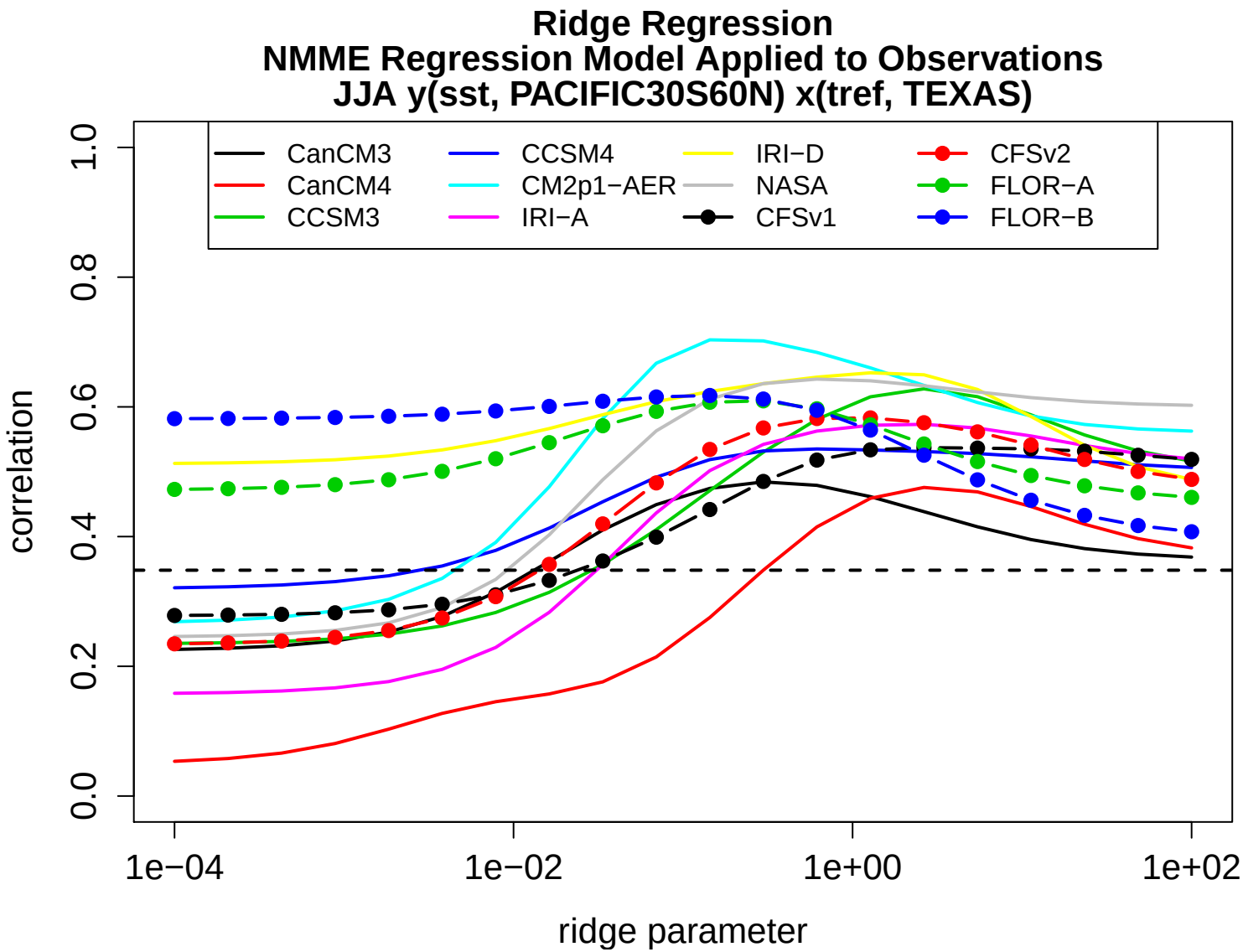
Principal Components Regression



LASSO


LASSO Prediction
Projection from Model to Observations
JJA y(sst, PACIFIC30S60N) x(tref, TEXAS)
CanCM3
CanCM4
CCSM3
CCSM4
CM2p1−AER
IRI−A
IRI−D
NASA
CFSv1
CFSv2
FLOR−A
FLOR−B
correlation
0.0
0.2
0.4
0.6
0.8
1.0
2e−04
1e−03
5e−03
2e−02
1e−01
5e−01
lasso parameter

Ridge


Ridge Regression
NMME Regression Model Applied to Observations
JJA y(sst, PACIFIC30S60N) x(tref, TEXAS)
CanCM3
CanCM4
CCSM3
CCSM4
CM2p1−AER
IRI−A
IRI−D
NASA
CFSv1
CFSv2
FLOR−A
FLOR−B
correlation
0.0
0.2
0.4
0.6
0.8
1.0
1e−04
1e−02
1e+00
1e+02
ridge parameter

Summary for Finding Relations in Climate Data

1. Data fishing is a serious problem.
2. Convincing evidence of relation involves showing:
   - relation holds in independent data
   - relation reproduced by climate models
   - relation understood from simple dynamical models
3. "Large-scale" principle can be framed as a sparsity problem.
4. N. Pacific is a major predictor of Texas temperature in models and observations.
5. LASSO and Ridge find relations that generalize to independent data better than EOF method.