Uncertainty in pattern scaling and addressing Big data

Douglas Nychka, National Center for Atmospheric Research

Summary

- Pattern scaling of climate model ensemble
- Nonstationary Gaussian fields
- Data analysis on super computers

Challenges:

Background

Emulating mean patterns and variability of temperature across and within scenarios with an application to RCPs 4.5 and 8.5. (2016). Stacey E Alexeeff, Stephan R. Sain, Doug Nychka, and Claudia Tebaldi

PART 1
Climate model emulation

, Charity,

Goals

Extend the value of a suite of climate model experiments using statistical models.

Substitute random processes for nonlinear chaotic processes.

Characterize the distribution of climate variables as a function of large scale processes.

Long term goal Use simple and aggregate variables to predict the distribution of climate for more complex variables and at finer scales.

Classic pattern scaling.

Patterns of temperature change over space are linear functions of the change in global mean temperature.

T_t Temperature at time t.

g_tGlobal mean temperature at time t.

$$(T_t - T_0) \approx P(g_t - g_0)$$

P is a slope that relates a change in global temperature to one locally.

If T_t is a temperature field then P is also interpreted as a field

$$(T_{i,t} - T_{i,0}) \approx P_i(g_t - g_0)$$

for i^{th} gridbox.

Emulation in general

Q_t Realization of climate field from a large model.

$$Q_t \sim \Gamma(\boldsymbol{h}_t, \boldsymbol{F}_t, \theta,)$$

The distribution of the field at t follows a probability distribution that depends of the (simpler process $\boldsymbol{h}_t$, external forcings $\boldsymbol{F}_t$, and some statistical parameters θ.

- A simpler process may eliminate the need for direct inclusion of forcings. (e.g. g_t based on an energy balance or intermediate climate model.)

- The distribution is much faster to compute than simulating additional cases of Q.

Context

- CESM Large Ensemble (CESM-LE), a 30-member initial condition ensemble of CESM simulations branched at 1930 under RCP 8.5
- CESM Medium Ensemble (CESM-ME), a 15-member initial condition ensemble for RCP4.5
- $\approx 1^{\circ}$ resolution, baseline period (1976 -2005), $t = 2006, \ldots, 2080$
- Seasonal averages of surface temperature

Parent study: Train on CESM-LE and predict CESM-ME

The emulator

A linear random effects model.

$$T_{i,t,k} - \bar{T}_{i,0} = a_{k,i} + \alpha_i + (b_{k,i} + \beta_i)(g_t - g_0) + \epsilon$$

RANDOM EFFECTS and FIXED EFFECTS

- Errors can be correlated in time
- Inclusion of intercept for emulation is open to interpretation.
- β_i is the P_i for pattern scaling

$(b_{k,i} + \beta_i)$ is estimated from ensemble (by OLS).

What is the uncertainty of β_i ?

What is the covariance of the random effects for the pattern?

PART 2
A spatial data problem

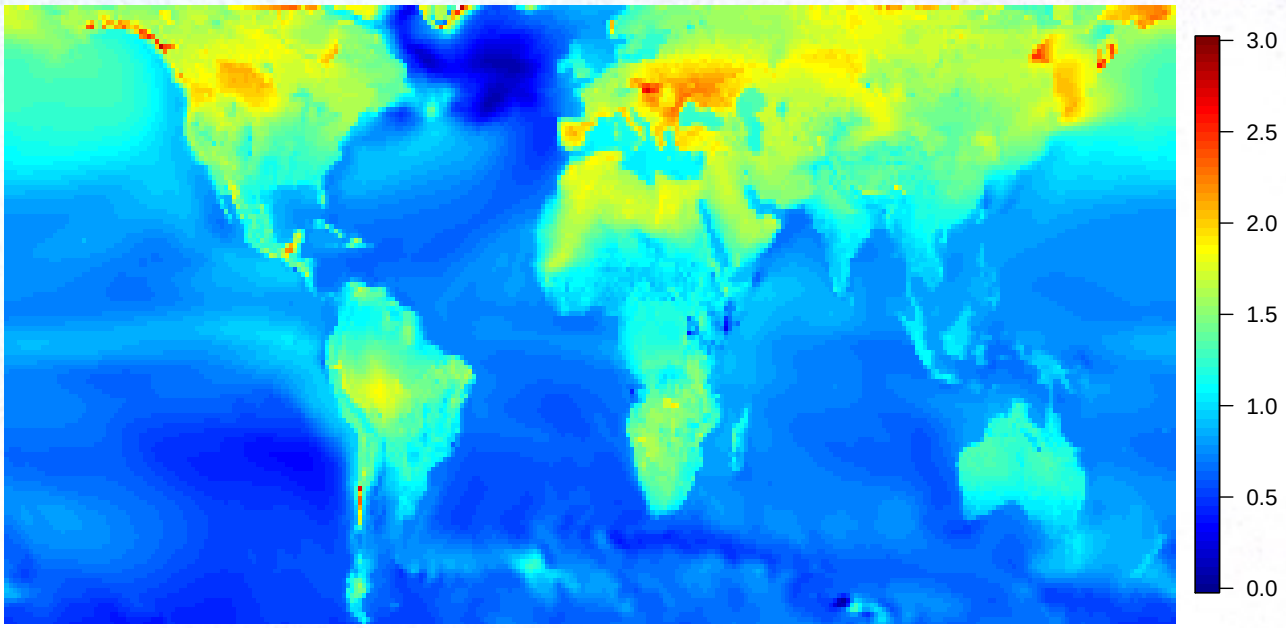
Hope, Charity, and Faith

Empirical slopes

$Y_{i,k}$ the slope from a simple linear regression of grid box i and ensemble member k on the global temperatures.

Y is are independent replicates of a spatial field.

Mean slopes across 30 members for JJA

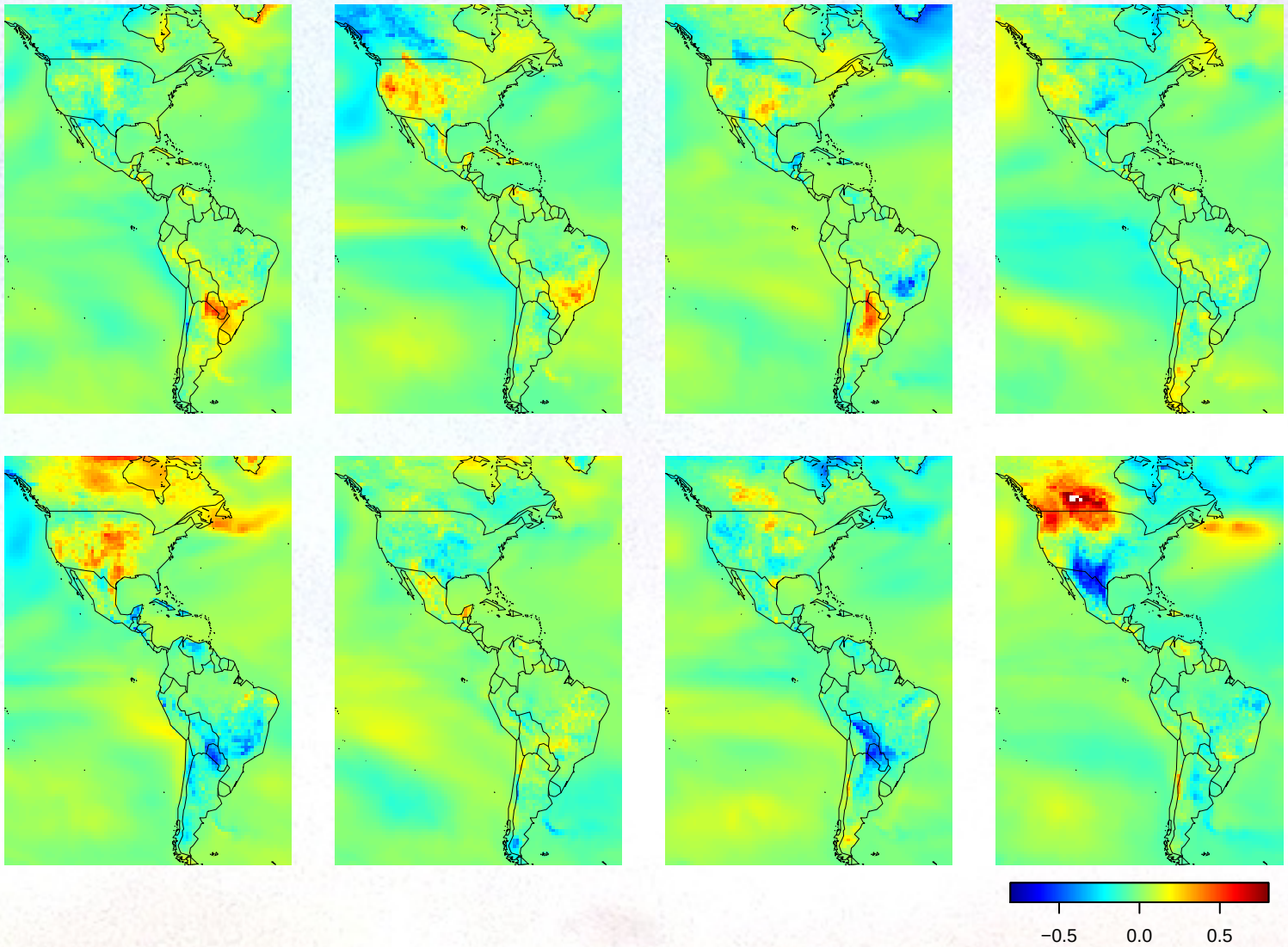


E.g value of 2.5 means: a 1° global increase implies 2.5° increase locally.

More on empirical slopes

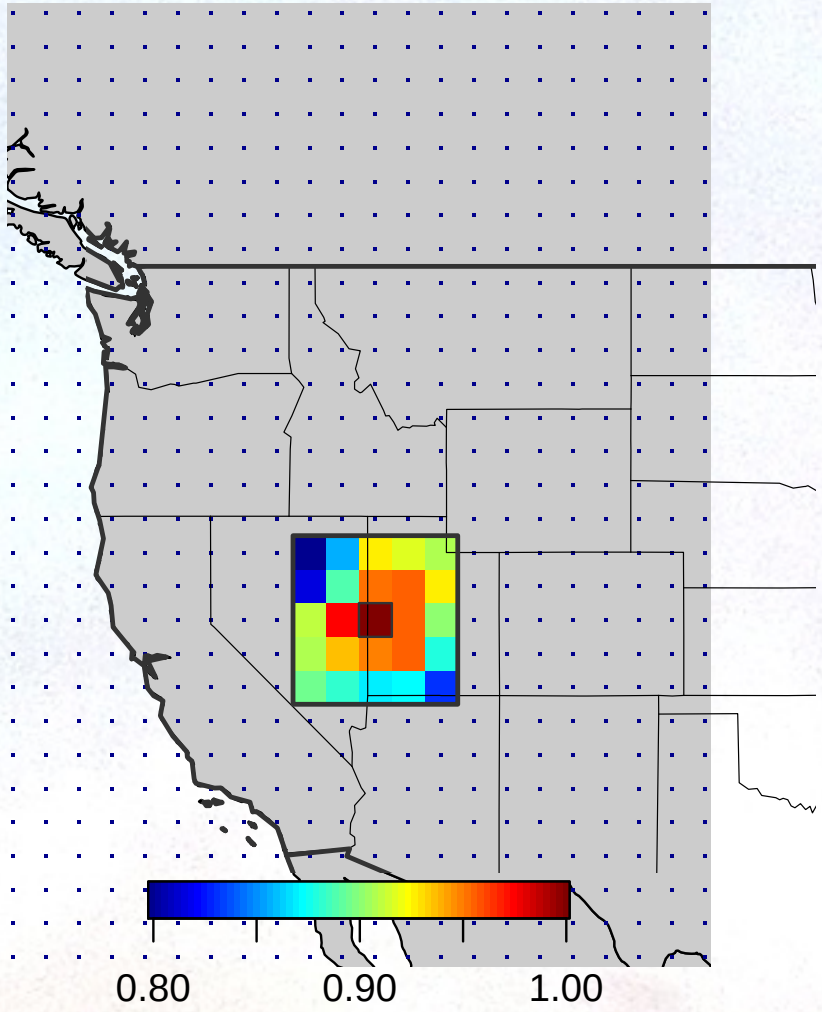
8 centered ensemble members

. . . there are 22 more of these!



A Local Spatial model

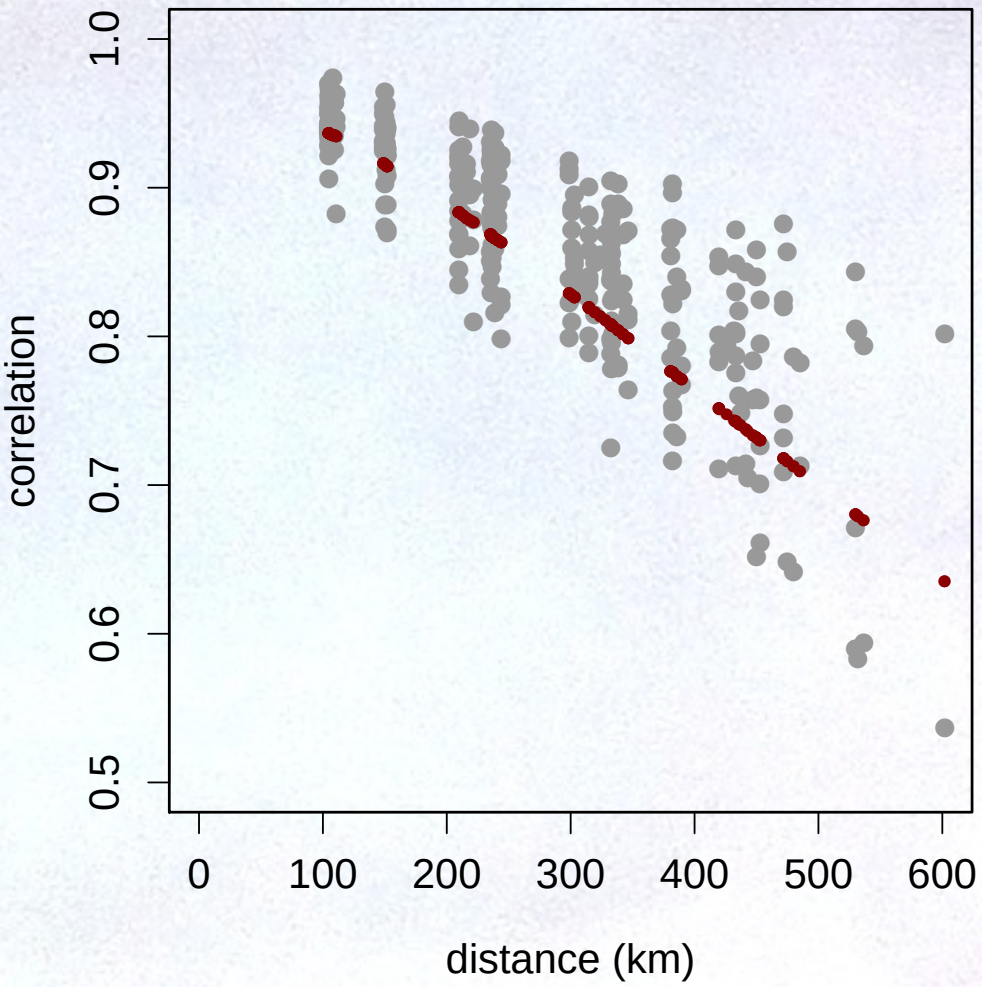
Correlations of ensemble members with center grid point. A 5×5 *grid.*



Fit a Matern spatial covariance function to these spatial data

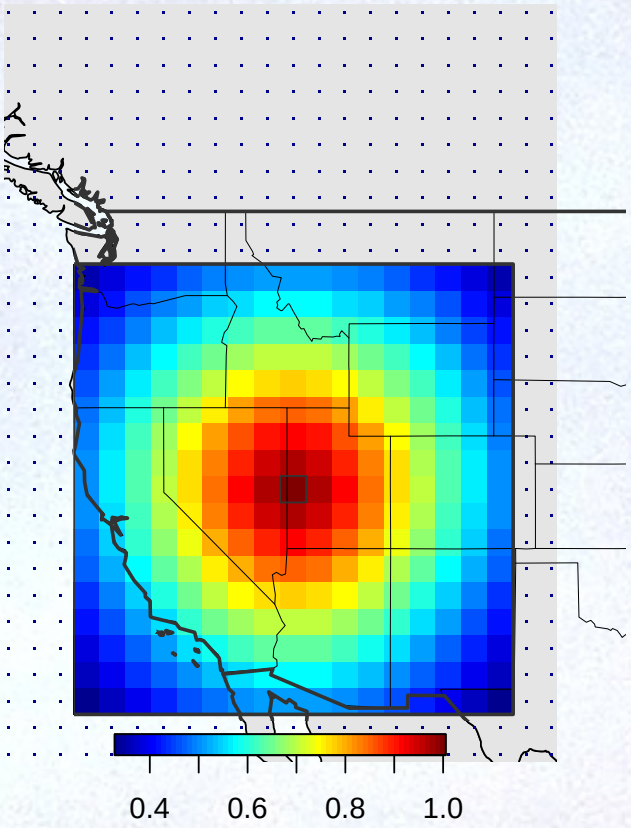
- Note: $5 \times 5 \times 30 = 750$ "data" points.
- Smoothness fixed at 1.0

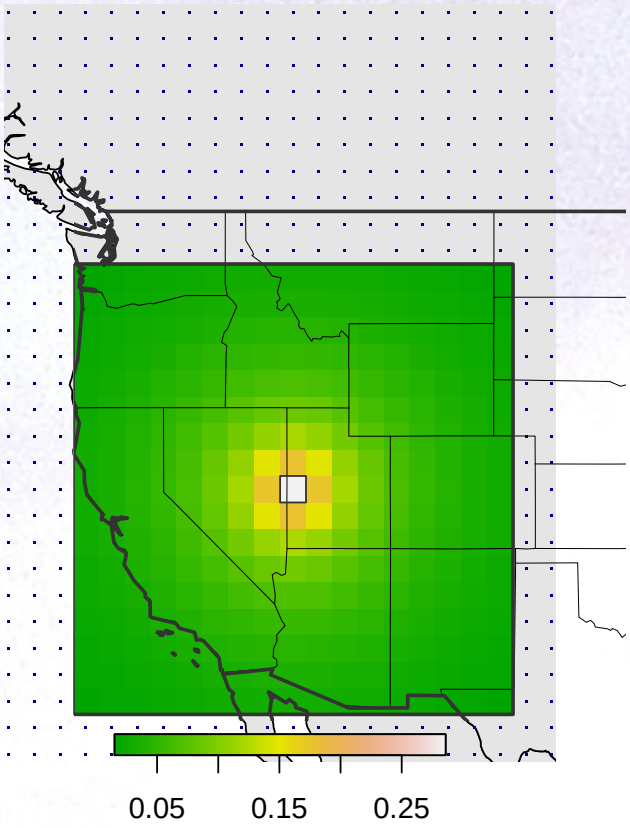
The Matern fit to correlations



Range parameter 431 km
Marginal standard deviation of process 0.175
Standard deviation of white noise component .033
Smoothness = 1.0

Covariance and weight function

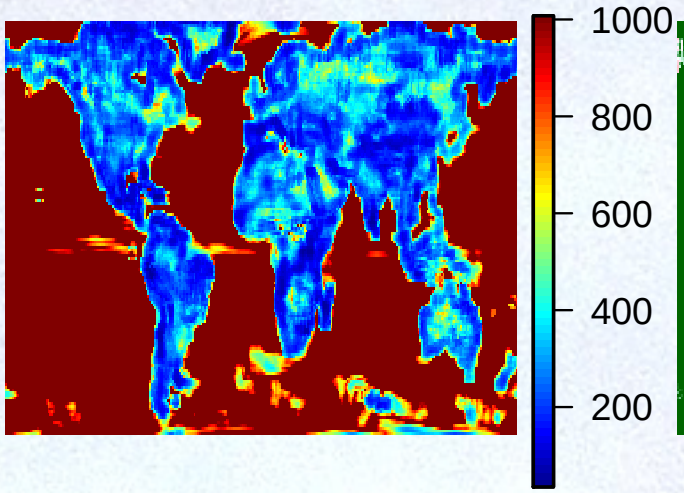




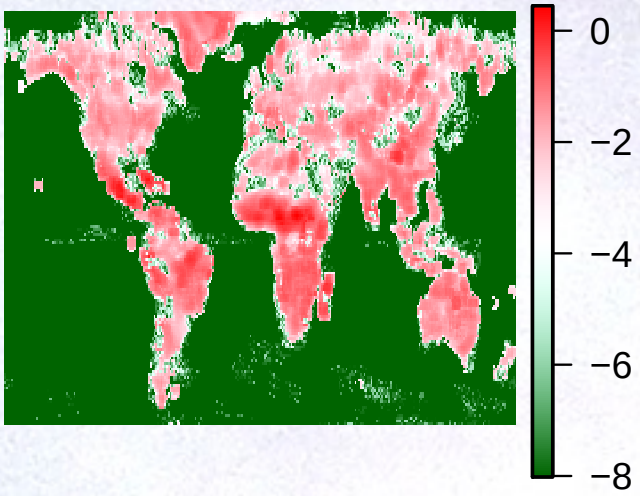
Recipe: Multiply independent N(0,1) at each grid point by the weight function. This is the value of the simulated field at the center grid point.

Fitting all grid boxes

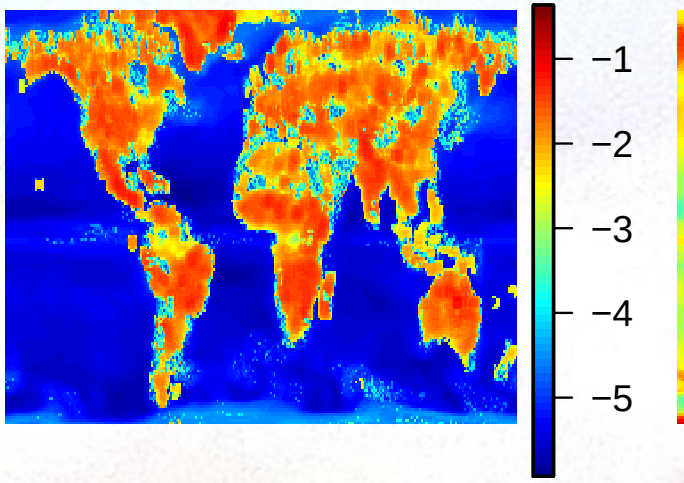
Correlation Range (km)



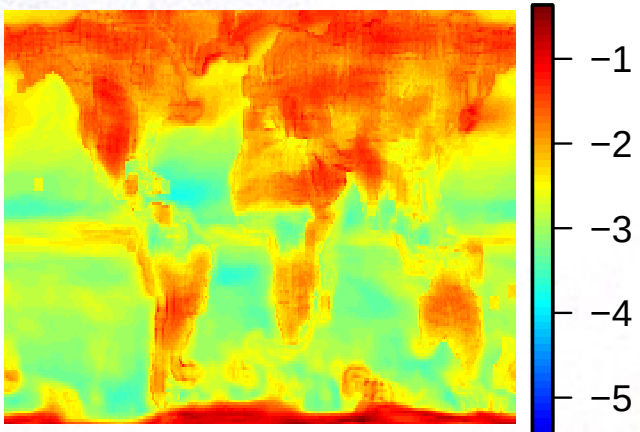
log10 lambda = sigma^2/ rho



log10 sigma

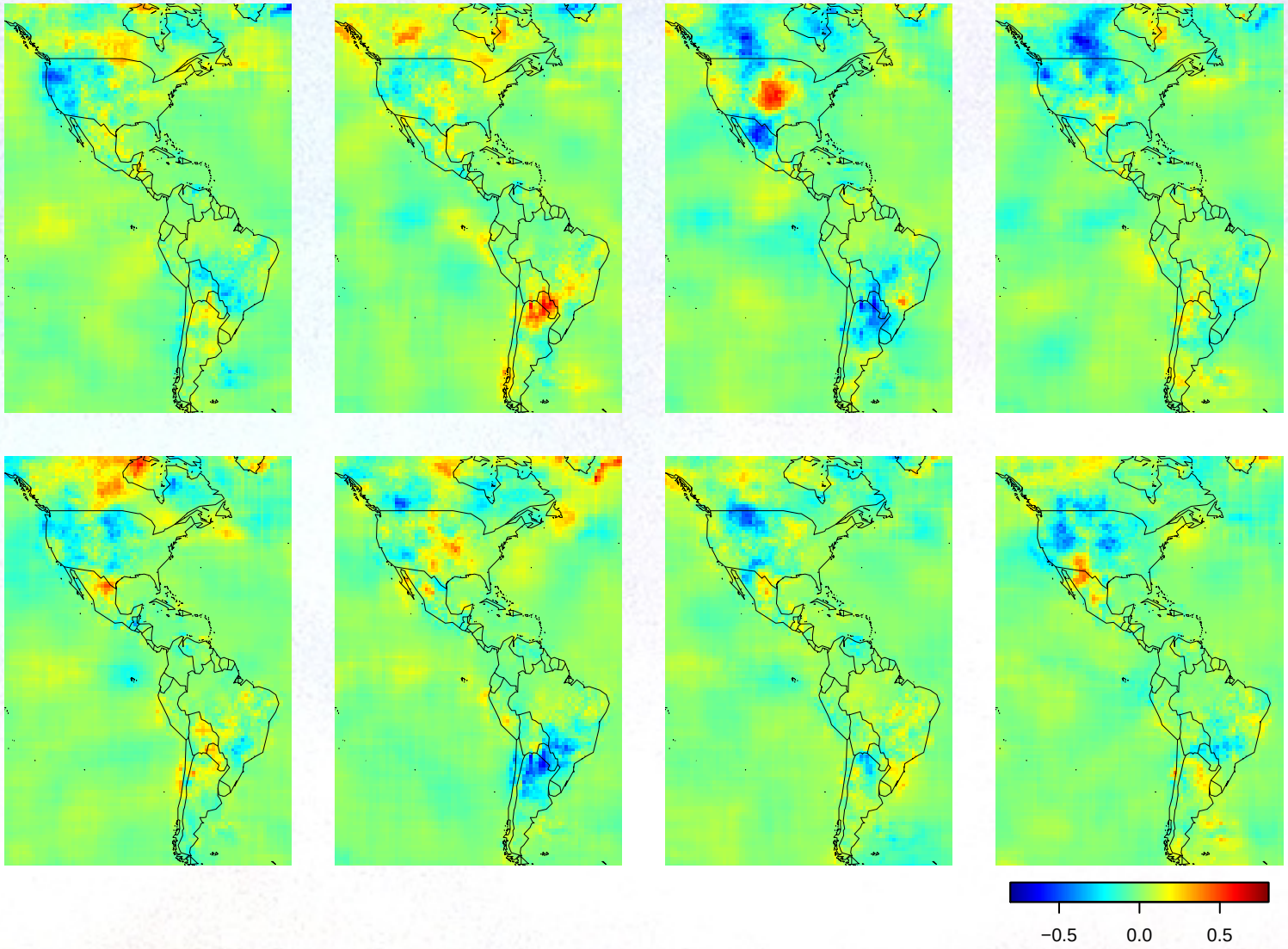


log10 rho



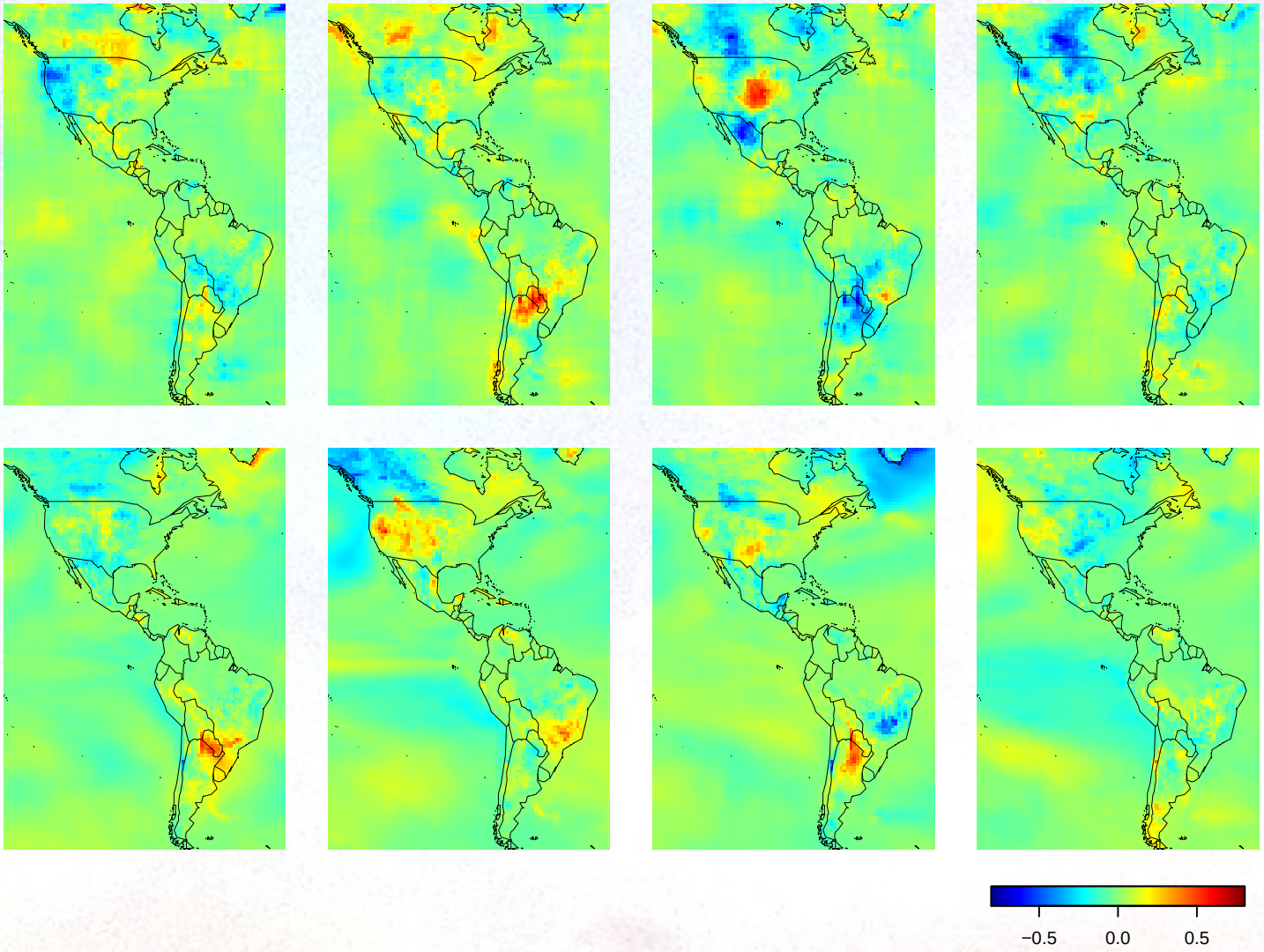
Does this work?

8 draws from spatial model.



Does this work?

4 draws from spatial model (top row).
4 ensemble members (bottom row)



PART 3:

Large spatial data sets

If I have to wait too long for my answer I forget my question.

– Rich Loft

Hope,

NCAR's supercomputer Yellowstone

≈72K cores = 4536 (nodes) × 16 (cores)
and each core with 2Gb memory
16 Pb parallel file system

- Core-hours are available to the NSF geosciences community with a friendly application process for student allocations.
- Accounts also available through collaboration with NCAR staff.

- *Supports R in both interactive and batch mode.*

The Supervisor R session.

In R ...

```
library(Rmpi)
# Spawn 256 workers
mpi.spawn.Rslaves(nslaves=256)
 # Broadcast the function to all workers

mpi.bcast.Robj2slave(fitCovariance)
 # apply this function to N tasks

output <- mpi.iapplyLB( 1:N, fitCovariance)


save( output, file="output.rda")
```

output is a list (N components) with the result for each case.

N $\approx$ 55K for 1° grid.

Computing for this project

Are many R worker processes feasible?

- `Rmpi` used to initiative many parallel, R sessions

from within a supervisor R session.

- Time to initiate 256 workers takes about 30-60 seconds
- Workers lose little time reading common data files – about 1 second.
- Time for output is about 2 seconds.

- Timing for main computation over 55K grid boxes

Cores	Window	Time (minutes)
Macbook (2)	5	$\approx$ 42
256	5	7.5
Macbook (2)	9	$\approx$ 320
16	9	40
64	9	10.3
256	9	7.5

Is a standard spatial analysis possible?

Embarrassing parallel steps:

Parameter estimation: Searching parameter space to maximize a likelihood or minimize cross validation mean square error.

Computing prediction error: Monte Carlo sampling from the error distribution (a.k.a. conditional simulation).

Iterate between spatial fitting and temporal fitting for space-time data

Iterate between local fitting and updating parameters

Summary

- Emulation of climate models for interpolation and uncertainty quantification is a good area for statistics.
- Local covariance fitting can capture variation in complex model output.
- There is a role for supercomputers to support interactive data analysis

Thank you!