

Nonstationary spatial data: think globally act locally

Douglas Nychka, National Center for Atmospheric Research



National Science Foundation

May 2018

Summary

- Two applications from climate science
- Nonstationary Gaussian fields
- Unconditional simulation of pattern scaling fields
- Conditional simulation of ocean temperature fields
- Big Data analysis on super computers (Big R)

Challenges:

Building convolution covariance models for large problems and actually computing the beasts!

Credits

- Pattern scaling - simulation, Ashton Weins (CU), Mitchell Crock (CU), Dorit Hammerling (NCAR).
- ARGO floats - conditional simulation, Mikael Kuusela (SAMSI), Michael Stein (UC-Rutgers), Pulong Ma (U Cincinnati)

Kuusela, M. and Stein M. (2017). Locally stationary spatio-temporal interpolation of Argo profiling float data
arXiv:1711.00460v2

PART 1

Spatial problems in climate science



Future Climate

What will the climate be in 60 years?

- Need a scenario of future human activities.

The representative concentration pathway (RCP) is a synthesis that specifies how greenhouse gases change over time.

- Need a geophysical model to relate the RCP to possible changes in climate.

Community Earth System Model (CESM)

A family of models developed at NCAR and supported by the National Science Foundation.

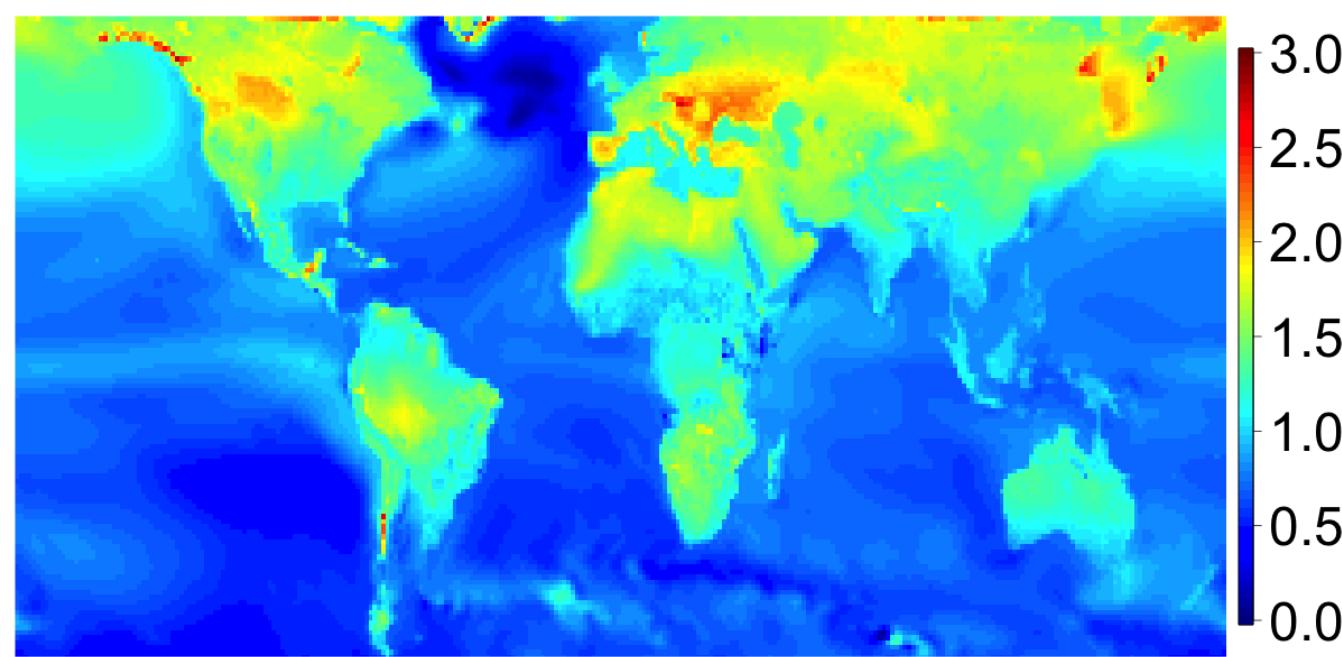
CESM Large Ensemble (CESM-LE)

A 30+ member ensemble of CESM simulations that have been designed
to study the local effects of climate change
– and the uncertainty due to the natural variability in the earth system.

- $\approx 1^\circ$ spatial resolution – about 55K locations
- Simulation period 1920 - 2080
- Using RCP 8.5 after 2005

Mean scaling pattern CESM

Slopes across 30 members for JJA

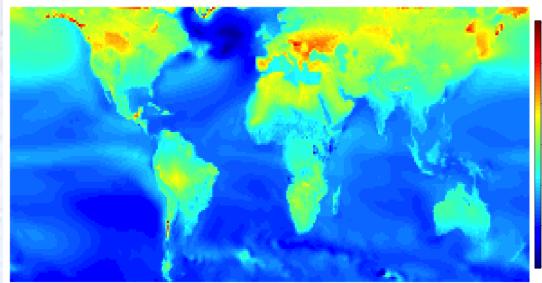


E. g. value of **2.5** means: a 1° global increase implies 2.5° increase locally.

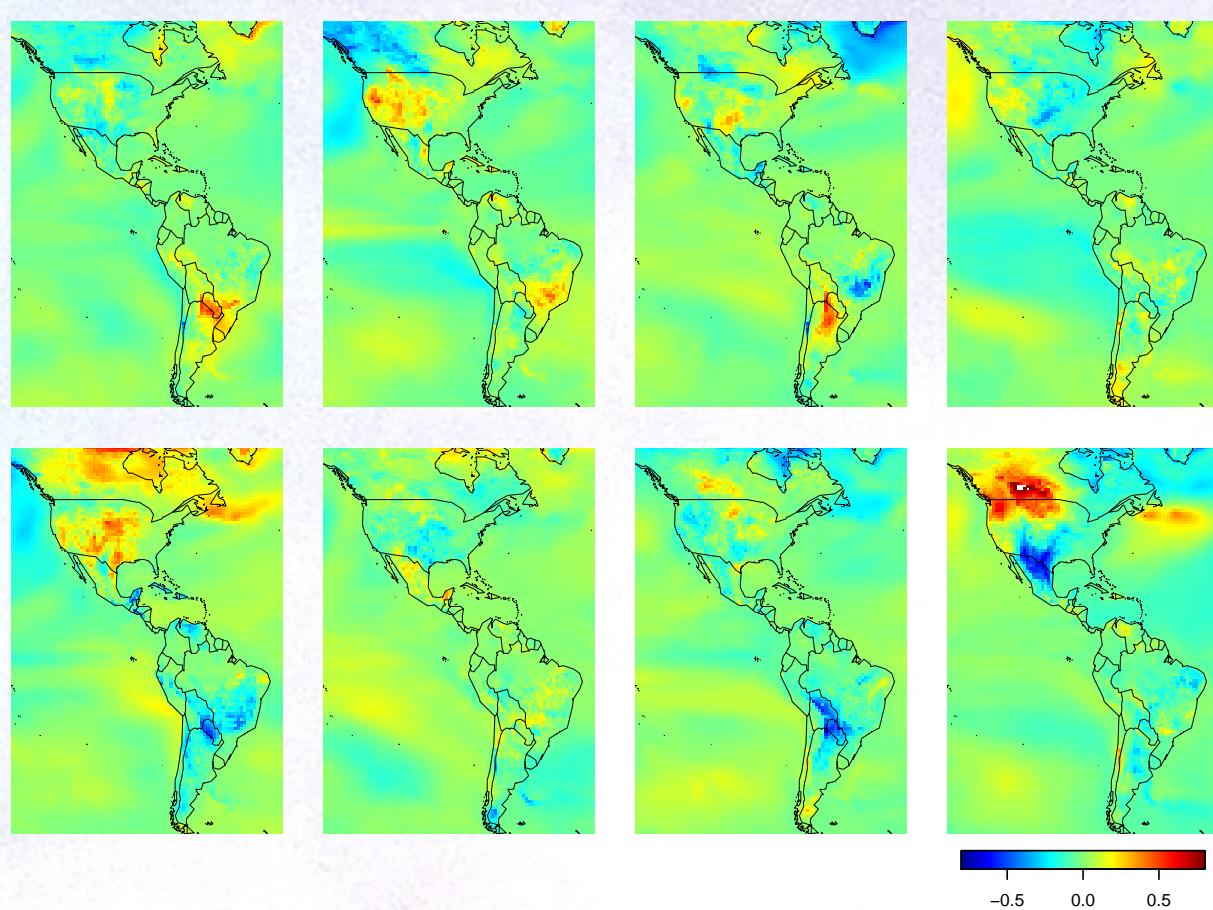
This allows us to determine the local mean temperature change based on a simpler model for the global average temperature

Individual patterns

Ensemble mean



First 8 out of 30 centered ensemble members



Goal: Simulate additional fields efficiently that match the spatial dependence in this 30 member ensemble.

PART 2

Nonstationary Gaussian Processes



Gaussian process models

$f(s)$ value of the field at location s .

$$E[f(s)] = 0 \text{ and } k(s_1, s_2) = E[f(s_1)f(s_2)]$$

- $f(s)$ is a Gaussian process if any finite collection of $\{f(s_1), \dots, f(s_N)\}$ has a multivariate normal distribution.
- f is mean square continuous (differentiable) if k is continuous (differentiable) in both s_1 and s_2 .
- An example of exponential covariance for a process that is stationary and isotropic:

$$k(s_1, s_2) = \sigma^2 e\left(\frac{\|s_1 - s_2\|}{\theta}\right).$$

- a strong assumption, note two covariance parameters σ and θ

Matérn covariance function

$$k(s_1, s_2) = \sigma^2 \text{Matern}_\nu(d) = \sigma^2 C d^\nu K_\nu(d),$$

and $d = \|s_1 - s_2\|/\theta$

- K_ν a modified Bessel function.
- C a normalizing constant depending on ν .
- Smoothness ν measures number of mean square derivatives and is equivalent to the polynomial tail behavior of the spectral density.
- σ^2 the process marginal variance
- When $\nu = .5$, Matérn is an exponential covariance, $\nu = \infty$, a Gaussian.

Nonstationary covariance functions

- Convolution model (Higdon, Fuentes)

Represent the process first, then figure out the covariance function

$$g(s) = \int_{\mathbb{R}^2} H(s, u) dW(u)$$

$dW(u)$ a two dimensional standard, white noise process.

The covariance function:

$$k(s_1, s_2) = \int_{\mathbb{R}^2} H(s_1, u) H(s_2, u) du$$

- H can be the Green's function for a stochastic PDE (– a connection to INLA)

2-D exponential kernel example:

$$H(\mathbf{s}, \mathbf{u}) = \frac{\sigma(\mathbf{s})}{\theta(\mathbf{s})} e^{-||\mathbf{s}-\mathbf{u}||/\theta(\mathbf{s})}$$

$$k_\theta(\mathbf{s}_1, \mathbf{s}_2) = \sigma(\mathbf{s}_1)\sigma(\mathbf{s}_2) \int \frac{1}{\theta(\mathbf{s}_1)\theta(\mathbf{s}_2)} e^{-||\mathbf{s}_1-\mathbf{u}||/\theta(\mathbf{s}_1)} e^{-||\mathbf{s}_2-\mathbf{u}||/\theta(\mathbf{s}_2)} d\mathbf{u}$$

- If $\theta(\mathbf{s}) \equiv \theta$ in 2-d this gives a Matérn with smoothness $\nu = 1.0$
- For unequal θ no simple closed form for this covariance.

Scale mixture (Paciorek, Stein)

$\nu = 1.0$

$$k(s_1, s_2) \sim d\mathcal{K}_1(d),$$

where

$$d = \frac{\|s_1 - s_2\|}{\sqrt{\theta(s_1)^2 + \theta(s_2)^2}}$$

These are different models.

Conjecture: as $\theta(s_2) \rightarrow 0$ give different smoothness at s_2

Open question how to figure out which model is more appropriate

Joint distribution

Observations

$$\mathbf{y}(s_i) = \mathbf{f}(s_i) + \mathbf{e}_i$$

$K_{i,j} = k(s_i, s_j)$ and \mathbf{f} is $MN(0, K)$

log Likelihood

$$\ell(\mathbf{y}, [\sigma^2, \theta, \tau]) = -(1/2)\mathbf{y}^T(K + \tau^2 I)^{-1}\mathbf{y} - (1/2)\log|K + \tau^2| + C$$

- Maximize to find parameters
- For large data sets K is also large.

Simulating a Gaussian Process

at locations s_1, \dots, s_M

- Form $K_{i,j} = k(s_i, s_j)$ covariance matrix at locations
- $f = \Omega e$ where e are iid $N(0, 1)$
 Ω is the matrix square root of K

Conditional simulation of a Gaussian Process

at locations s_1^g, \dots, s_M^g

conditional on observations y_1, \dots, y_N at s_1^o, \dots, s_N^o

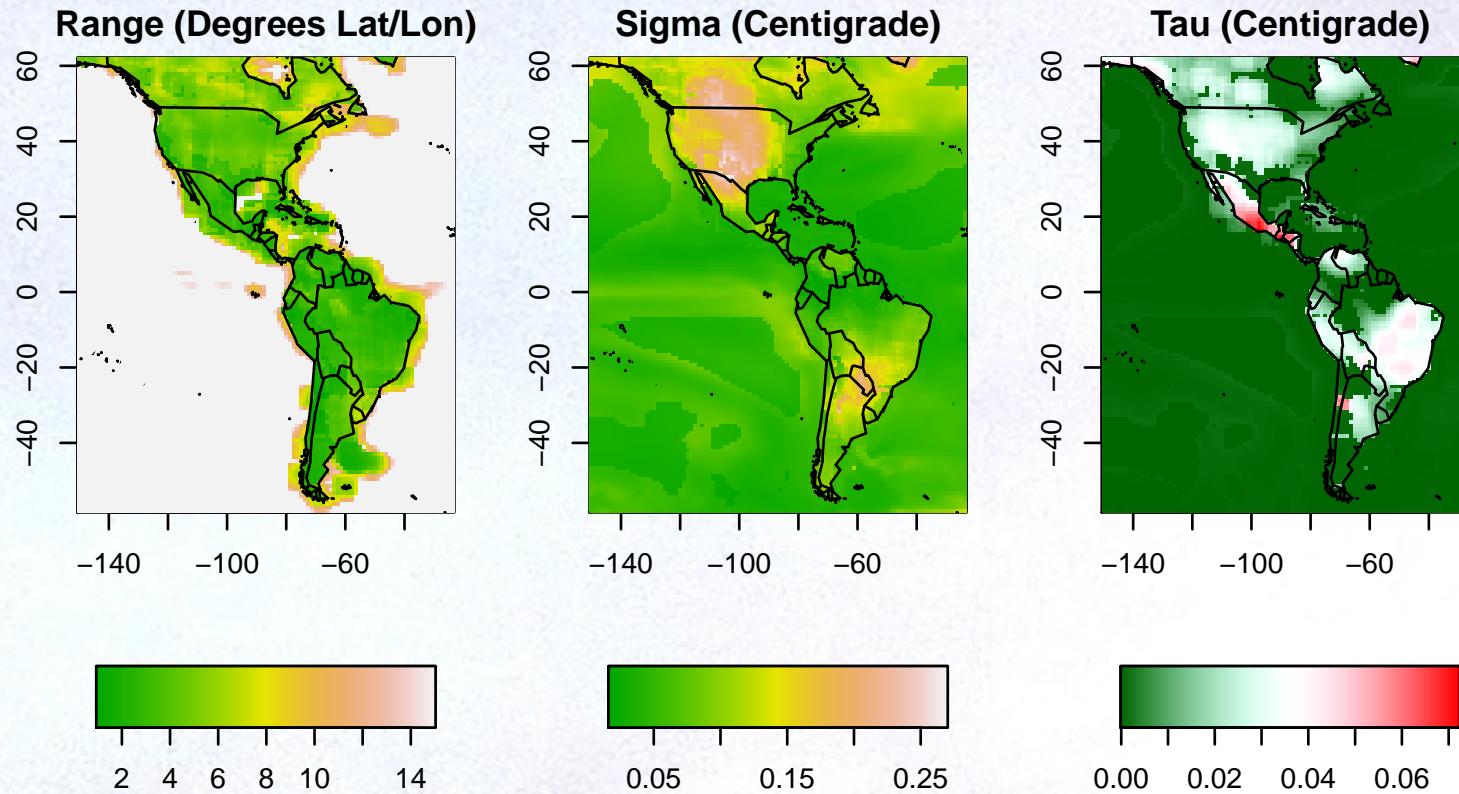
- Form $K_{o,o}$ Covariance matrix at observations locations
 $K_{g,g}$ Covariance matrix at grid locations
 $K_{g,o}$ Cross-covariance matrix between grid and observation locations
- $f = \hat{f} + \Omega e$ where e are iid $N(0, 1)$
 \hat{f} the conditional expectation for bbf (aka Kriging)
 Ω is the matrix square root of $K_{g,g} - K_{g,o}(K_{o,o} + \tau^2 I)^{-1}K_{g,o}^T$

PART 3 Unconditional simulation



Climate model patterns

Local Matérn MLEs for the 30 member ensemble patterns



- 11× 11 windows using coordinates in degrees
- About 13K grid boxes in this subregion

What should we do with these?

- Assume that the parameter estimates at the center of the window are good estimates for the parameter “fields” $\sigma(s)$, $\theta(s)$, and $\tau(s)$.
- RECALL Form $K_{i,j} = k(s_i, s_j)$ covariance matrix at all observation locations
- $f = \Omega e$ where e are iid $N(0, 1)$
 Ω is the matrix square root of K

PROBLEM:

K is too big for computation.

SOLUTION:

Reexpress model in more computable form.

- Approximate the nonstationary model with a spatial autoregressive model (SAR)
(Parameters of the local Matérn models encoded as parameter fields in the LatticeKrig model.)
- Exploit sparse matrix methods to implement
 $f = \Omega e$ where e are iid $N(0, 1)$
 Ω^T is the sparse Cholesky decomposition of K

A Spatial Autoregression (SAR)

Gridded field:

$$\begin{matrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & c_1 & \cdot & \cdot \\ \cdot & c_2 & c_* & c_3 & \cdot \\ \cdot & \cdot & c_4 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{matrix}$$

SAR weights:

$$\begin{matrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot \\ \cdot & -1 & a(s) & -1 & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{matrix}$$

The filter:

$$a(s)c_* - (c_1 + c_2 + c_3 + c_4) = \text{white noise}$$

- $a(s)$ needs to be greater than 4.
 $1/\sqrt{a(s) - 4}$ – an approximate range parameter
- $B\mathbf{c} = \text{i.i.d.} N(0, 1)$ where B is a sparse matrix
- Covariance for \mathbf{c} is $(B^T B)^{-1} = Q^{-1} = K$

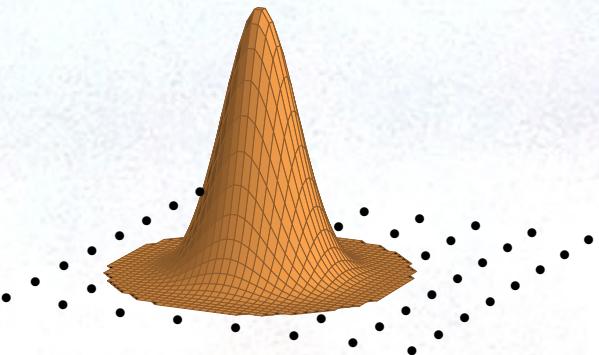
Representing a random surface

$$g(x) = \sum_j \phi_j(x) c_j$$

- c is the random field from the SAR.
- $\{\phi_j(x)\}$ are compact, radial basis functions :

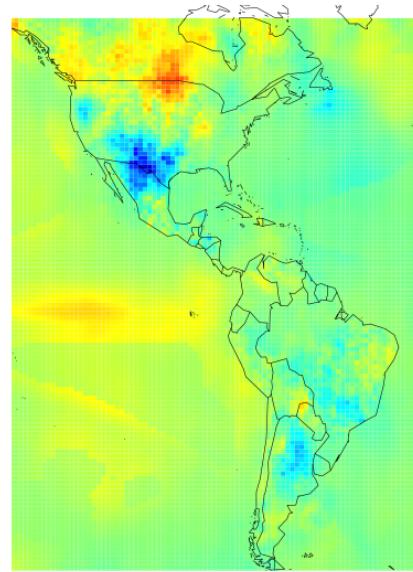
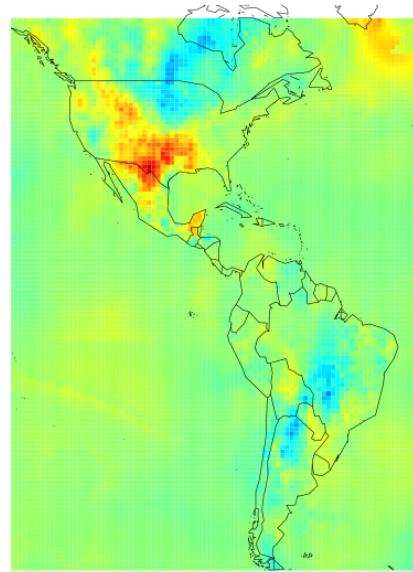
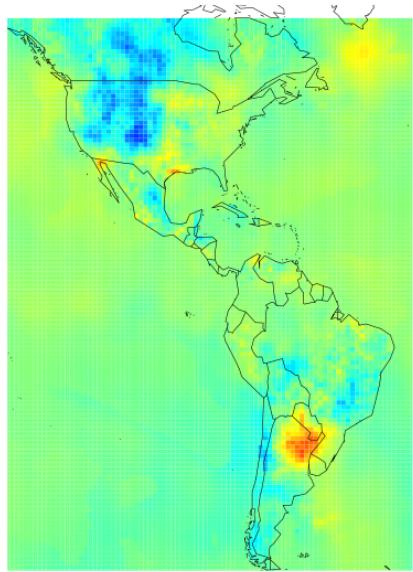
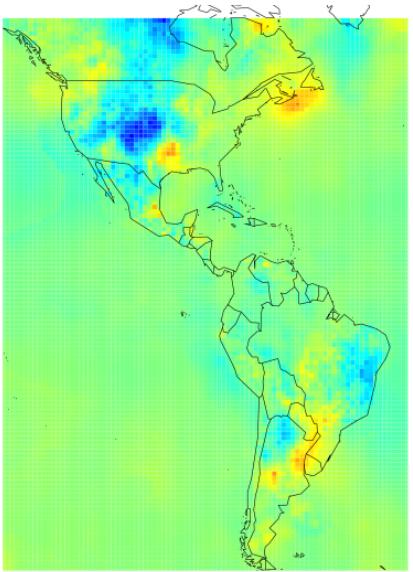
$$\phi_j(x) = \psi(||s - u_j||/\delta)$$

A member of the Wendland basis functions

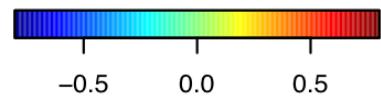
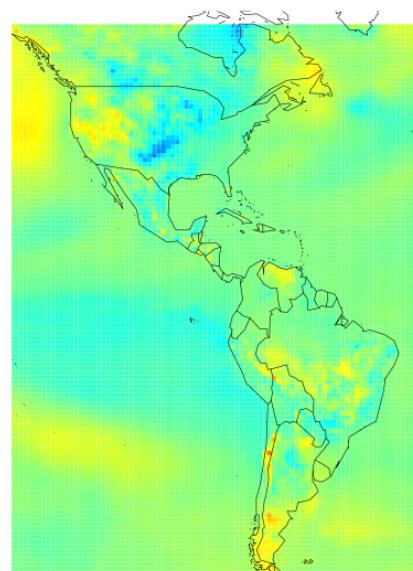
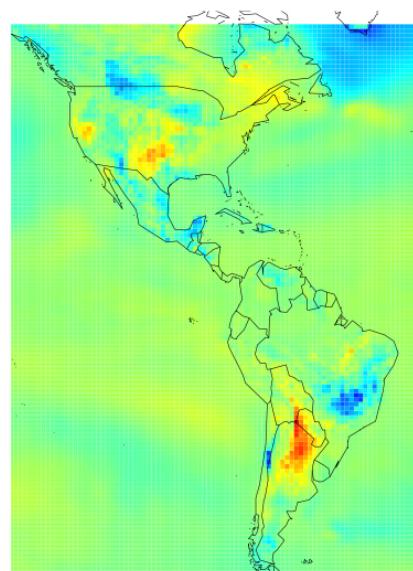
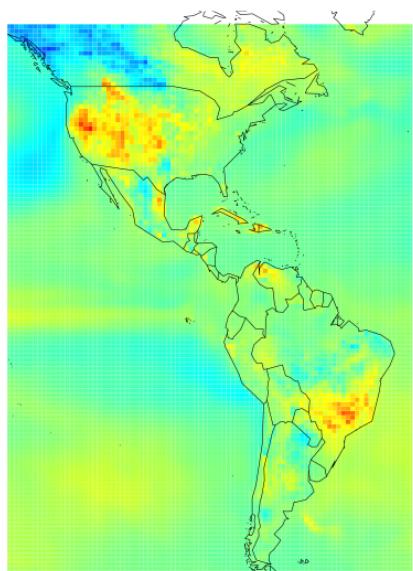
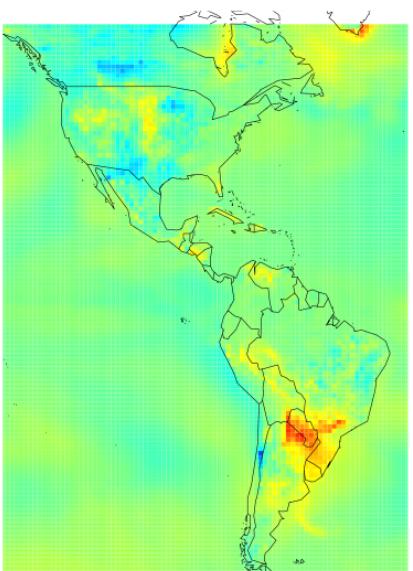


Emulating pattern scaling fields

Statistical model



Ensemble members



PART 4

Conditional simulation



What should we do with these?

- Assume that the parameter estimates at the center of the window are good estimates for the parameter “fields” $\sigma(s)$, $\theta(s)$, and $\tau(s)$.
- RECALL
- $f = \hat{f} + \Omega e$ where e are iid $N(0, 1)$
 \hat{f} the conditional expectation for bbf (aka Kriging)
 Ω is the matrix square root of

$$K_{g,g} - K_{g,o}(K_{o,o})^{-1}K_{g,o}^T$$

PROBLEM: all the K s are too big for computation.

SOLUTION: Simulate conditional field by moving local neighborhoods

- Generate a realization of e on the grid.

LOOP OVER GRID LOCATIONS

- For each grid location evaluate Ω in a local neighbor centered at this point, Ω_{local}
- Find the symmetric square root of Ω_{local}
- Apply the center row of square root matrix to the right subset of e .
(throw the other rows away!)

END LOOP

This is an embarrassingly parallel computation.

PART 5:

Parallel computation with R



The Cheyenne supercomputer.



$\approx 145K$ cores = 4032 nodes \times 36 cores
and each core with 2Gb memory
52Pb parallel file system

- Core-hours are available to the NSF research community.
- Simple application process for graduate student allocations.
- Implementation of R on batch and interactive nodes.

Are zillions of R workers feasible?



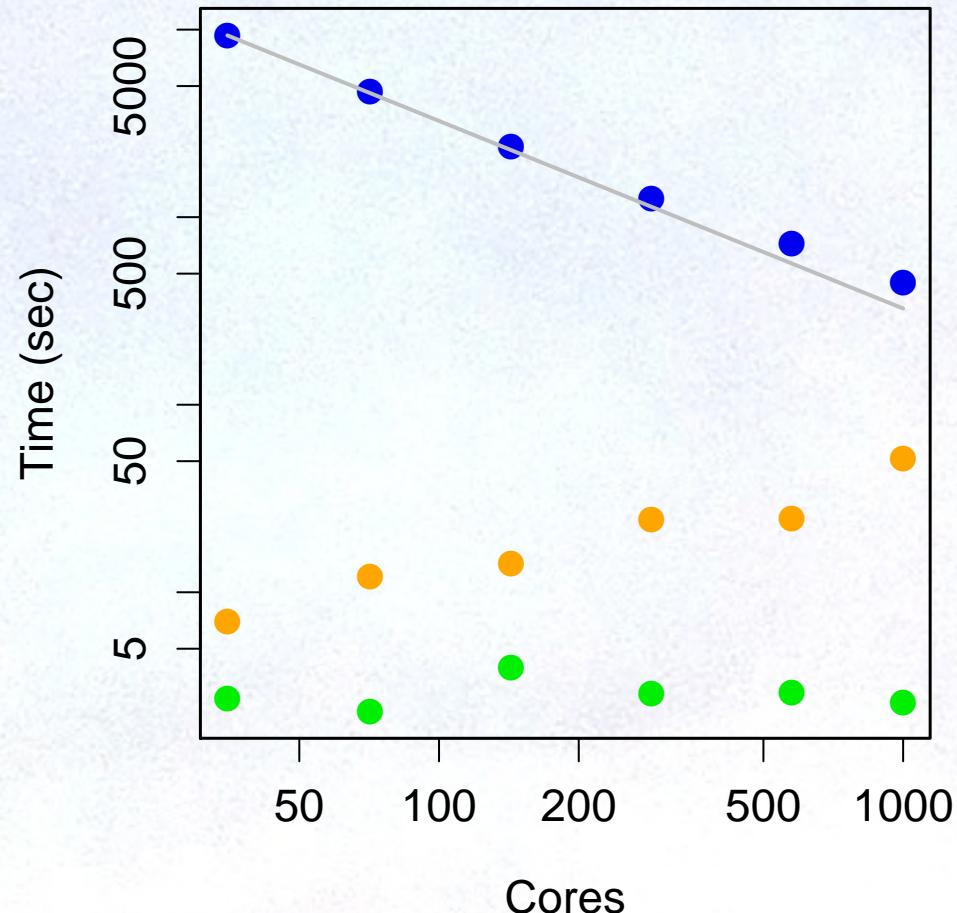
Yes for embarrassingly parallel data analysis.

- `Rmpi` used to initiate many parallel R sessions from within a supervisor R session.
- Time to initiate 1000 workers takes about 1 minute.
- Little time lost in broadcasting the data object (12Mb) – about 3 seconds.

Approximate linear scaling using Rmpi

Individual times for:

spawn broadcast apply



Wall clock time in seconds to fit 1000 9×9 blocks with the LatticeKrig model.

Summary

- Emulation of climate model experiments for interpolation and uncertainty quantification is a fruitful area for data science.
- Local covariance fitting can capture variation in complex model output and in geophysical fields.
- Markov random field based models are suited for large data sets.
- There is an emerging role for supercomputers to support data analysis.

Software

- `fields` R package, Nychka et al. (2000 - present)
- `LatticeKrig` R package, Nychka et al. (2014- present)
- HPC4Stats SAMSI short course August 2017, Nychka, Hammerling and Lenssen.

Background reading

Nychka,D., Hammerling, D. , Krock, M. Wiens, A. (2017). Modeling and emulation of nonstationary Gaussian fields.
arXiv:1711.08077

Kuusela , M and Stein M. (2017). Locally stationary spatio-temporal interpolation of Argo profiling float data
arXiv:1711.00460v2

Alexeeff, S. E., Nychka, D., Sain, S. R., & Tebaldi, C. (2016). Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments.
Climatic Change, 1-15.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., & Sain, S. (2015). A multi-resolution Gaussian process model for the analysis of large spatial datasets.

Thank you!

