Uncertainty in the pattern scaling of climate models

Douglas Nychka, National Center for Atmospheric Research
Summary

- Pattern scaling of a climate model ensemble (Climate Science)
- Nonstationary Gaussian fields (Applied Stats)
- Spatial autoregressions, multiscale processes (Big Data)
- Results for CESM ensemble
- Data analysis on super computers (Big R )

Challenges:
Building covariance models for large problems, simulating Gaussian fields, and actually computing the beasts!
PART 1
Climate model emulation
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.

Earth system models are complex, physical models that integrate the feedbacks in the atmosphere, ocean, land, ice and other components to determine the climate under different conditions.

**Community Earth System Model (CESM)**
A family of models developed at NCAR and supported by the National Science Foundation.
Main benefit of a climate model is in quantifying the local effects of climate change.

**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
Classic pattern scaling.

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

\[ T_t \text{ Temperature at time } t \text{ at a specific location (or climate model grid box)} \]

\[ g_t \text{ Global mean temperature at time } t. \]

\[ (T_t - T_0) \approx P(g_t - g_0) \]

\( P \) is a slope (amplitude) that relates a change in global temperature to one locally.
Mean 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.
PART 2
A spatial data problem
Individual patterns

Ensemble mean

First 8 centered ensemble members

... there are 22 more of these!
**Goal:**
Be able to simulate additional fields efficiently that match the spatial dependence in this 30 member ensemble.

**Result:**
• Use simple models for other RCP scenarios to generate a large number of ensembles and without the need to do full CESM runs.

• Enable researchers using climate change impact models to drive their models with a larger number of ensembles and without the need to do full CESM runs.
Gaussian process (GP) models

\( f(x) \) value of the field at location \( x \).

\[ E[f(x)] = 0 \text{ and } k(x_1, x_2) = E[f(x_1)f(x_2)] \]

- \( f(x) \) is a GP if any finite collection of \( \{f(x_1), \ldots, f(x_N)\} \) has a multivariate normal distribution.

- \( f \) is mean square continuous (differentiable) if \( k \) is continuous (differentiable) in both \( x_1 \) and \( x_2 \).

- If the process is stationary and isotropic then

\[ k(x_1, x_2) = \varphi(||x_1 - x_2||). \]

- a strong assumption but identifies the direct connection with Radial Basis Functions methods
Estimating covariance parameters by maximum likelihood.

Covariance now depends on some parameters, $k_\theta$.

Log Likelihood:

$$\ell(f, \theta) = \frac{1}{2} f^T (K_\theta^{-1}) f - \frac{1}{2} \ln |K_\theta| + \text{constants}$$

- $\hat{\theta} =$ maximizer of $\ell(f, \theta)$ – aka the MLE
- Inference (e.g. confidence intervals) based on the curvature of the log likelihood surface at $\hat{\theta}$

To simulate a GP at locations $x_1, \ldots, x_M$

- Form $K_{\theta,i,j} = k_\theta(x_i, x_j)$
- $f = K_\theta^{1/2} e$ where $e$ are iid $N(0,1)$
The Matérn covariance

\[ k(x_1, x_2) = \rho C d^\nu \mathcal{K}_\nu(d), \text{ and } d = ||x_1 - x_2|| / \alpha \]

- \( \mathcal{K}_\nu \) a modified Bessel function.
- \( C \) a normalizing constant depending on \( \nu \).
- Smoothness \( \nu \) measures number of mean square derivatives and is equivalent to the polynomial tail behavior of the spectral density.
- When \( \nu = .5 \), Matérn is an exponential covariance, \( \nu = \infty \), a Gaussian.
Challenges with large spatial data

- $K$ a dense covariance matrix that has the dimensions of the number of locations. (e.g. 55K for CESM output)
- For statistical computations it is not feasible to evaluate $K^{1/2}$, $K^{-1}$ and $|K|$ when the data set is large.
- Typically the GP is not stationary – so a heroic global computation may not solve our problem!

Solution
Fit local GP models to a moving window of the data.
A local spatial model

Correlations of center grid point.

A 5 \times 5 grid of output, 25 \times 30 data points.

Matérn fit with MLEs and \nu = 1.0

Data model:
\[ P(x_i) = f(x_i) + \epsilon_i \]
\( f \) – Matérn GP
\( \epsilon_i \) – uncorrelated noise
Fitting all grid boxes

Matérn MLEs

- 11×11 windows using coordinates in degrees
- About 13K grid boxes in this subregion
More on nonstationary covariances

- Convolution model (Higdon, Fuentes)
  with exponential kernels;

\[
k_\theta(x_1, x_2) \sim \int e^{-\|x_1 - u\|/\theta(x_1)} e^{-\|u - x_2\|/\theta(x_2)} du
\]

If \(\theta(x) \equiv \theta\) in 2-d this gives a Matern with smoothness \(\nu = 1.0\)

- Scale mixture (Paciorek, Stein)
  \(\nu = 1.0\)

\[
k(x_1, x_2) \sim dK_1(d),
\]

where

\[
d = \frac{\|x_1 - x_2\|}{\sqrt{\theta(x_1)^2 + \theta(x_2)^2}}
\]

These are different models.

**Conjecture:** as \(\theta(x_2) \rightarrow 0\) give different smoothness at \(x_2\)
Some drawbacks

- Approximate simulation does not address long range dependence beyond the local neighborhoods.
- This is not a global model and the covariance is implicit.
- Simulation is slow because each grid point requires a separate eigen-decomposition.

*These points motivate a different approach to representing the GP*
PART 3:

Markov random fields, multi-scale processes
A Spatial Autoregression (SAR)

Gridded field:

\[
\begin{array}{cccccccc}
& & & & & & & \\
& & & & & & & \\
& & & & C_1 & & & \\
& & & C_2 & C_* & C_3 & & \\
& & & C_4 & & & & \\
& & & & & & & \\
\end{array}
\]

SAR weights:

\[
\begin{array}{cccccccc}
& & & & & & & \\
& & & & & & & \\
& & & -1 & & & & \\
& & -1 & a & -1 & & & \\
& & -1 & & & & & \\
& & & & & & & \\
\end{array}
\]

The filter:

\[
aC_* - (C_1 + C_2 + C_3 + C_4) = \text{white noise}
\]

- \( a \) needs to be greater than 4.
- \( Bc = \text{i.i.d.} N(0,1) \) where \( B \) is a sparse matrix
- Covariance for \( c \) is \( (B^T B)^{-1} = Q^{-1} \).

\( Q \) known as the precision or information matrix.
Generalization ...

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 radial basis functions:

\[ \phi_j(x) = \psi(||x - u_j||/\delta) \]

A member of the Wendland basis functions
Multi-resolution extension

Add together different resolution levels to approximate the Matérn GP.

\[ g(x) = \sum_j \phi_j(x)c_j + \sum_j \phi_j(x)c_j + \ldots + \sum_j \phi_j(x)c_j \]

\[ g(x) = g_1(x) + g_2(x) + \ldots + g_L(x) \]

Correlation functions for \( g_1, \ldots, g_6 \) and a target exponential
Weighting by asymptotic theory for exponential models

\[ \text{VAR}(g_{\text{level}}) = 2^{-\text{level}/2} \]

that is \( \approx 0.707, 0.50, 0.354, 0.25, 0.177, 0.125 \).

Correlation functions adding levels and the target exponential model.
Computing

To simulate the GP
Basis function matrix: \( \Phi_{i,j} = \phi_j(x_i) \)

- Sparse solve of \( V^T c = e \) where \( e \) are iid \( N(0, 1) \)
  \( V \) the Choleski factor of \( Q \), i.e. \( Q = V^T V \)

- \( f = \Phi c \).

- This a global simulation for a well-defined GP.

- Includes possible long-range correlations.

- At least two orders of magnitude faster than the local approach.
PART 4:
Nonstationary Analysis
A global model:
Can encode all rows of $B$ from local fitting giving the sparse precision matrix $Q = BB^T$

A lazy (but faster) route
Fit the local Matérn model and approximate using the LatticeKrig process model. From $\hat{\theta}, \hat{\nu}$ translate to weights and $\alpha$.

Future Analysis
Estimate the weights and $\alpha$ directly for the LatticeKrig model.
Emulating model fields

Statistical model

Ensemble members

NCAR D. Nychka Pattern Scaling

National Science Foundation
The land/ocean boundary

Now work directly with the LatticeKrig process parameters

Land/ Ocean mask over California coast

\[ \alpha = -1 \text{ isotropic same as the ocean links} \]
\[ \alpha = 0 \text{ no correlation of grid box with land} \]
Boundary parameters

Using a 9X9 moving window and maximum likelihood
PART 5:
Parallel computation with R
The Cheyenne supercomputer.

≈145K cores = 4032 nodes × 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 embarassingly 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 $9 \times 9$ 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.
- 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

arXiv:1711.08077


Thank you!