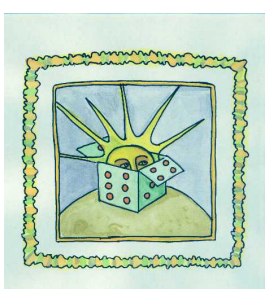
Models for models

Douglas Nychka Geophysical Statistics Project
National Center for Atmospheric Research

Outline

- Statistical models and tools
- Spatial fields (Wavelets)
- Climate regimes (Regression and clustering)
- Subgrid scale features (Markov models)



Supported by the National Science Foundation DMS

Overview

"Data"

Numerical models produce data at different scales. Statistical techniques can be use to productively summarize and probe model output.

Statistical models and tools

Perhaps the most important foundation for statistics is the use of statistical models to derive tools for data analysis, inference and representation.

The past 20 years have witnessed a rapid increase in flexible models for complicated processes and features.

An example based on the curve fitting problem

Suppose f is a "smooth" 1-d curve and we observe

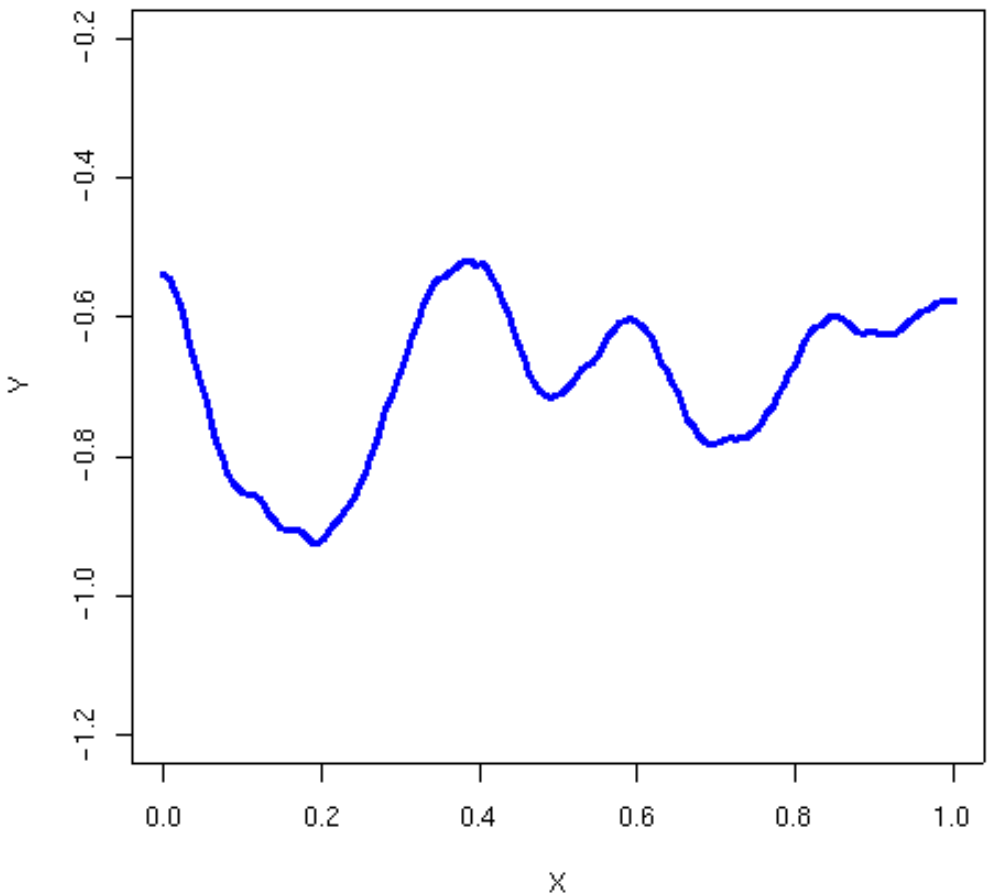
$$Y_k = f(x_k) + noise$$

Find f!

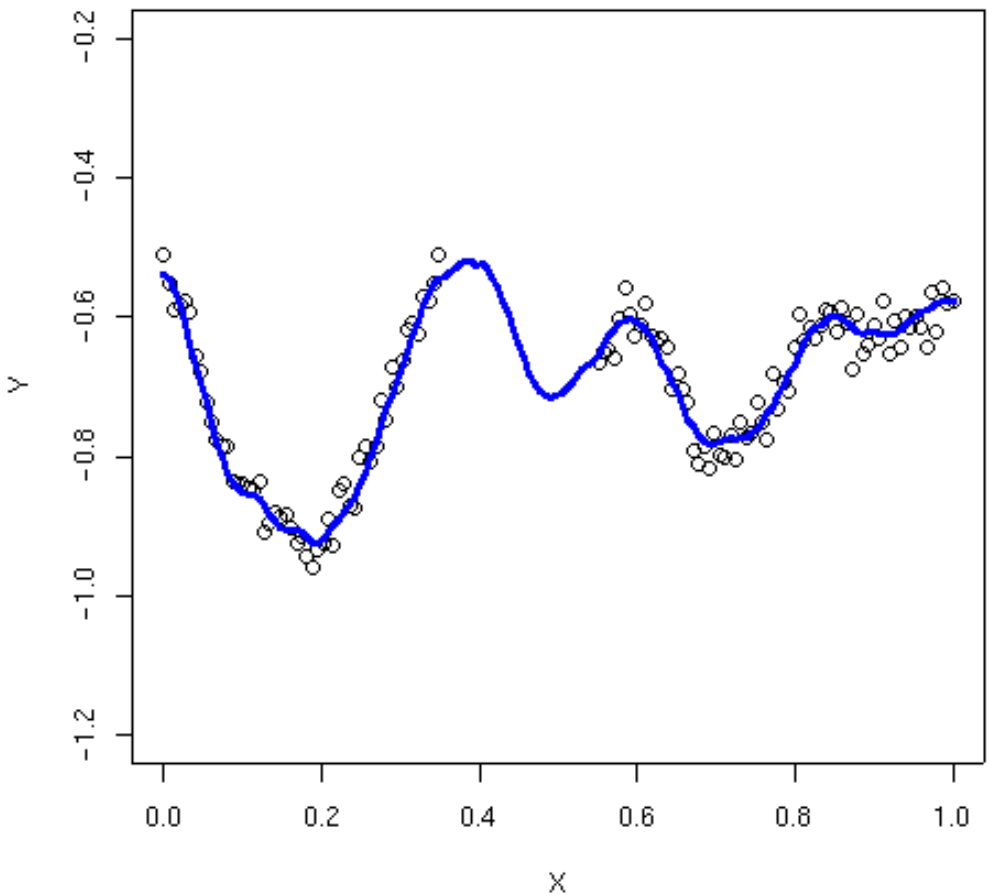
Some formalism:

- *Prior knowledge on f*
  $[f]$ A stochastic model for f, e.g. f is a smooth Gaussian process with an Matern covariance (ν=1.5, θ=1).
  or f is a member of a function class (continuous)
- *Link the data to f*
  $[Y|f]$ distribution of the data given knowledge of f
- *Combine information*
  $[f|Y]$ distribution of f informed by the data.
  or estimate f so it has good performance "on the average"

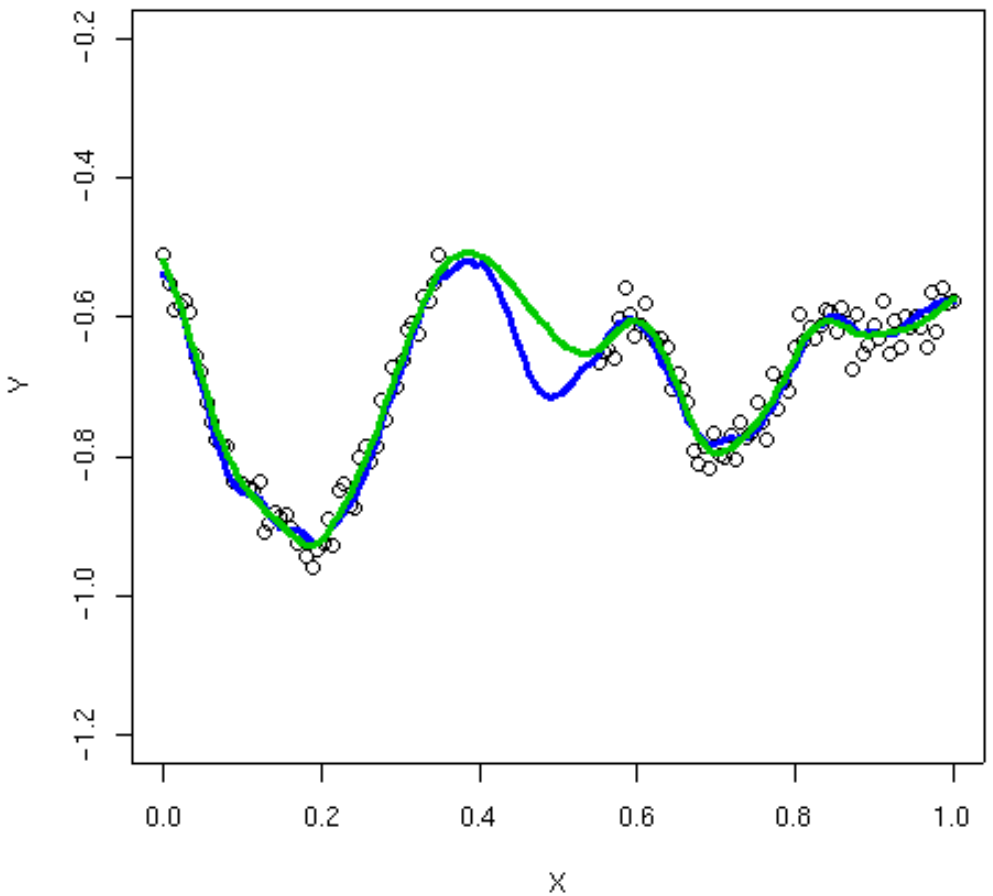
”True function”



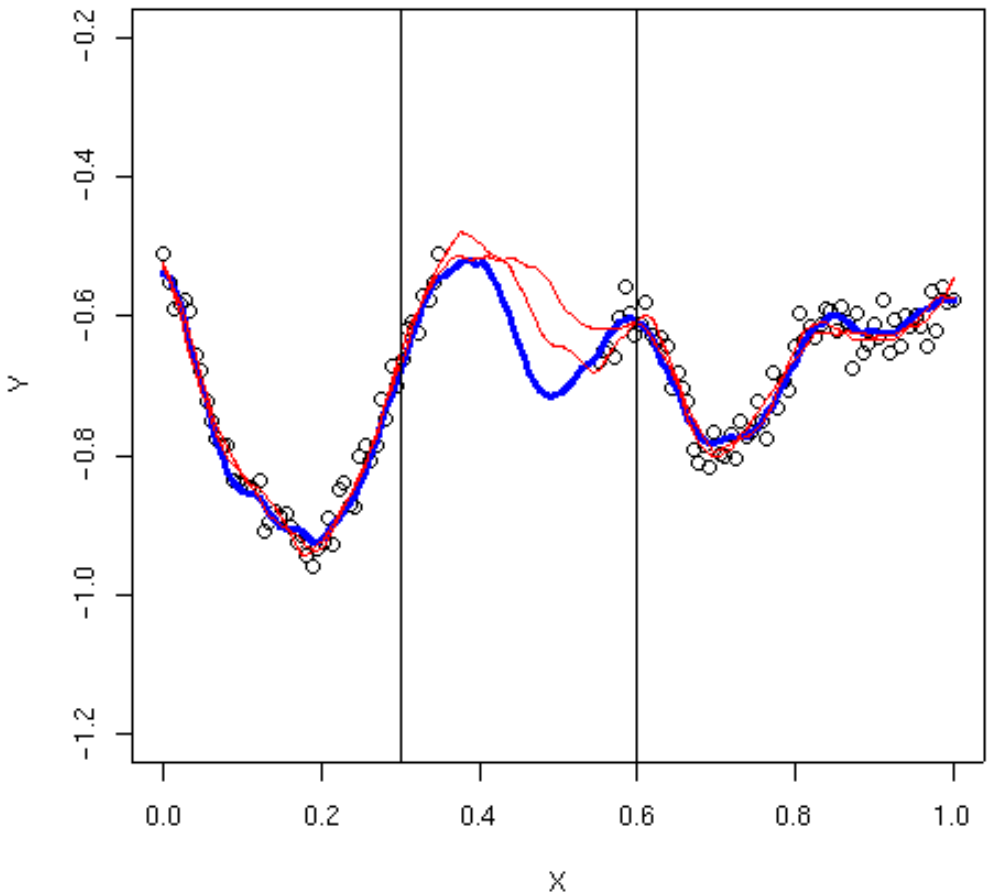
"True function" with noisy observations



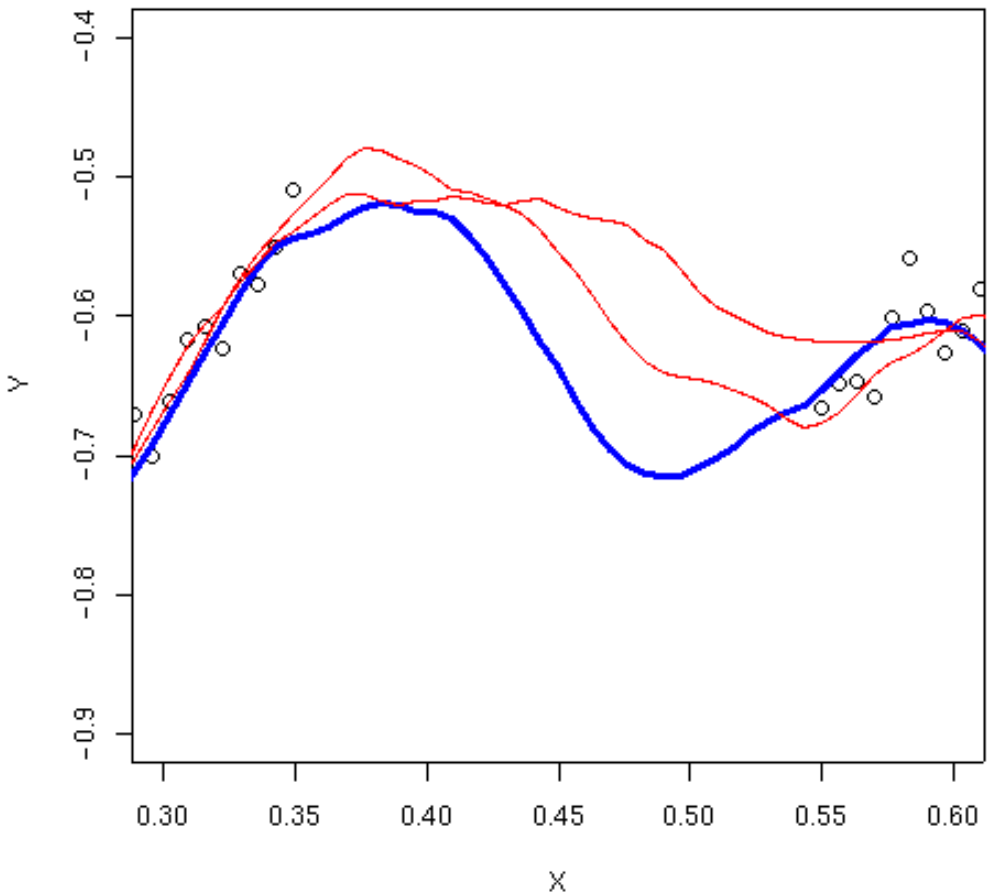
"True function" with smooth estimate



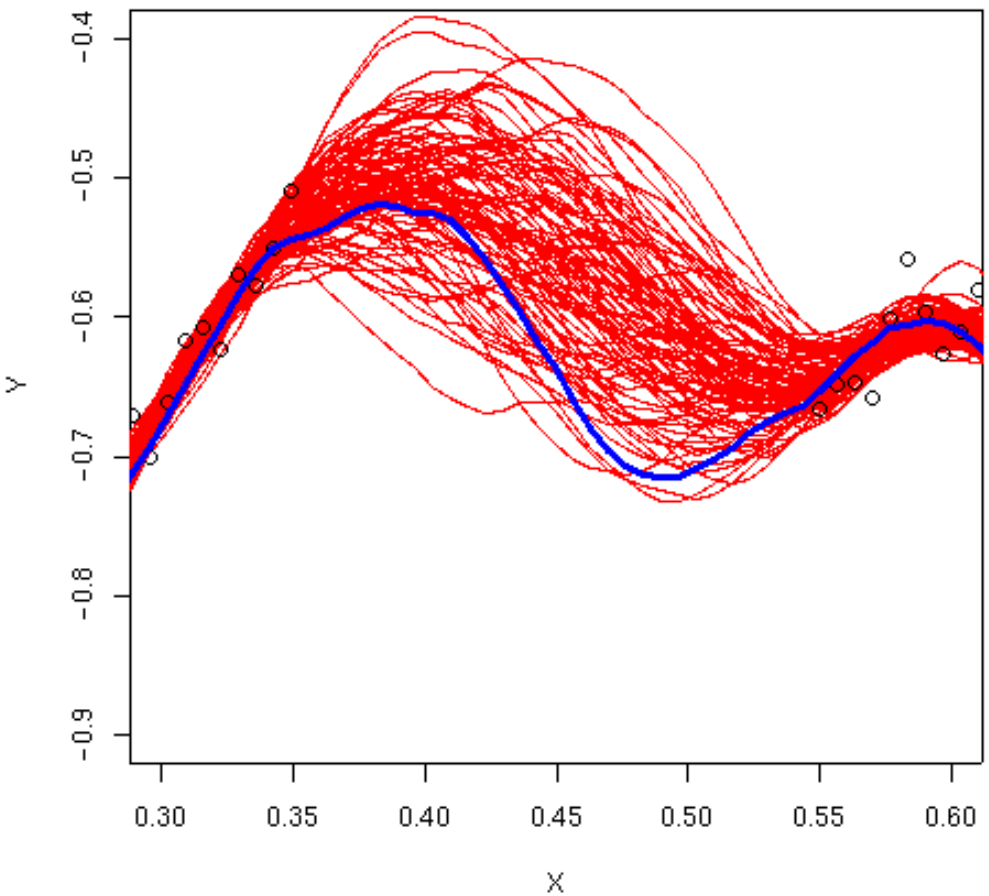
"True function" with two draws from $[f|y]$, the posterior



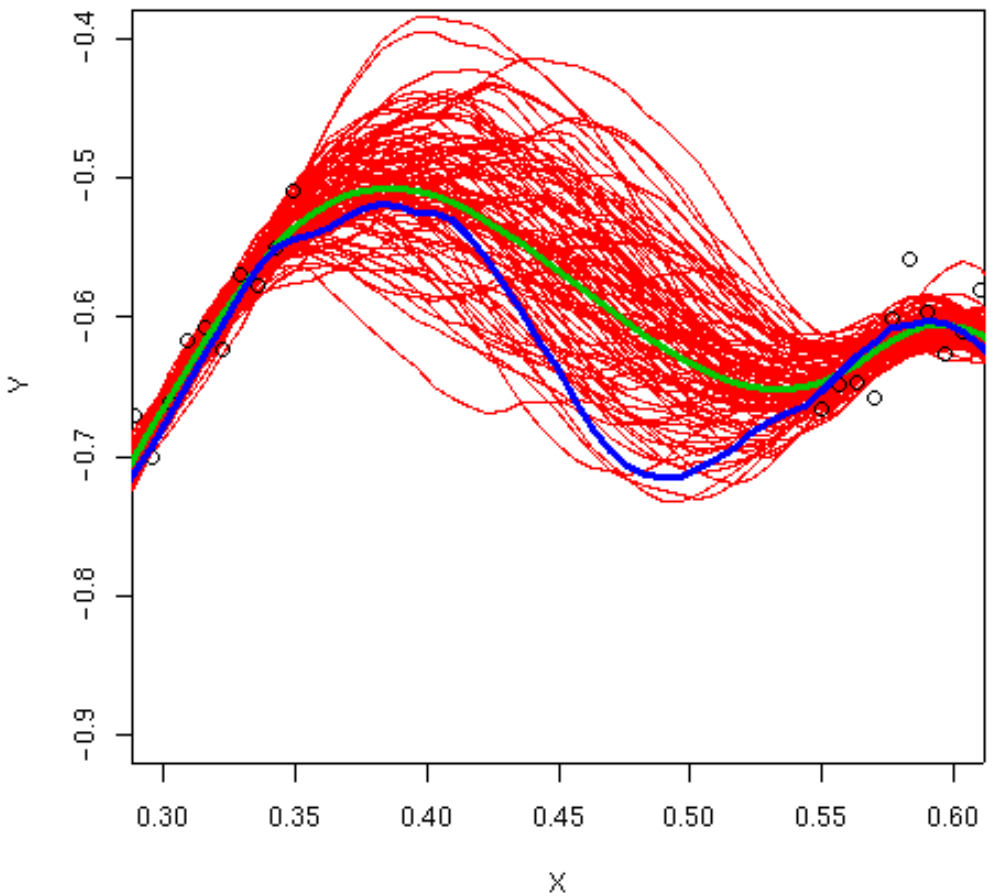
"True function" with two draws from $[f|y]$, the posterior



"True function" with 100 draws from the posterior



100 draws from the posterior and the mean



Some Remarks

A key ingredient here was the choice of model for f. Here we simplified things to a choice of covariance function assuming a Gaussian process.

If a Bayes approach is not taken, the uncertainty characterization is possible, but more difficult.

The model for $[f]$ can be hierarchical and span many space/time scales.
(See Berliner, Wikle presentations and also work of Cressie and coworkers)

The model of $[Y|f]$ can include multiscale data types!

Covariance models

J. Andrew Royle, Chris Wikle, William Cox

EPA Regional Oxidant Model

Model output:

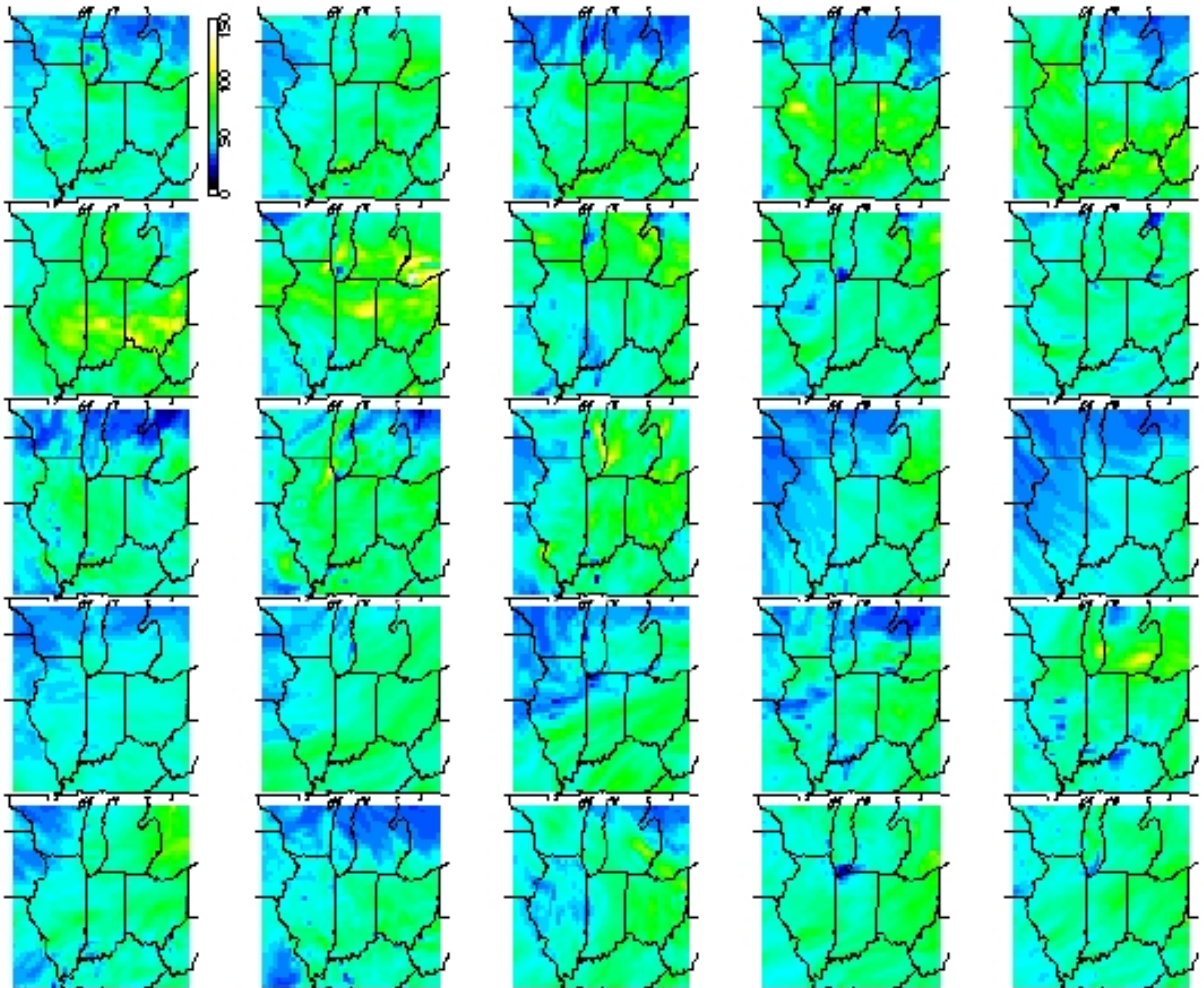
8 hour daily average ozone

48×48 grid centered on Illinois and Ohio,

grid box size: 25km

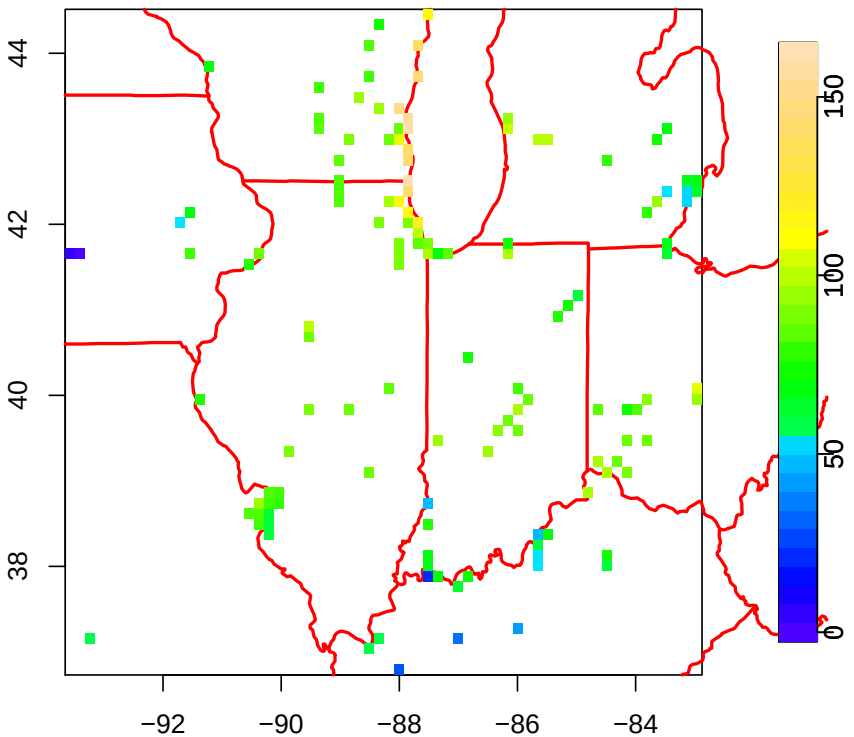
79 days in June-August 1987, this was a period of high summer ozone

Sequence of 25 days (11-35)



One goal is to estimate a covariance model that can be used for spatial prediction and analysis of stations data.

Daily ozone for 6/18/1987

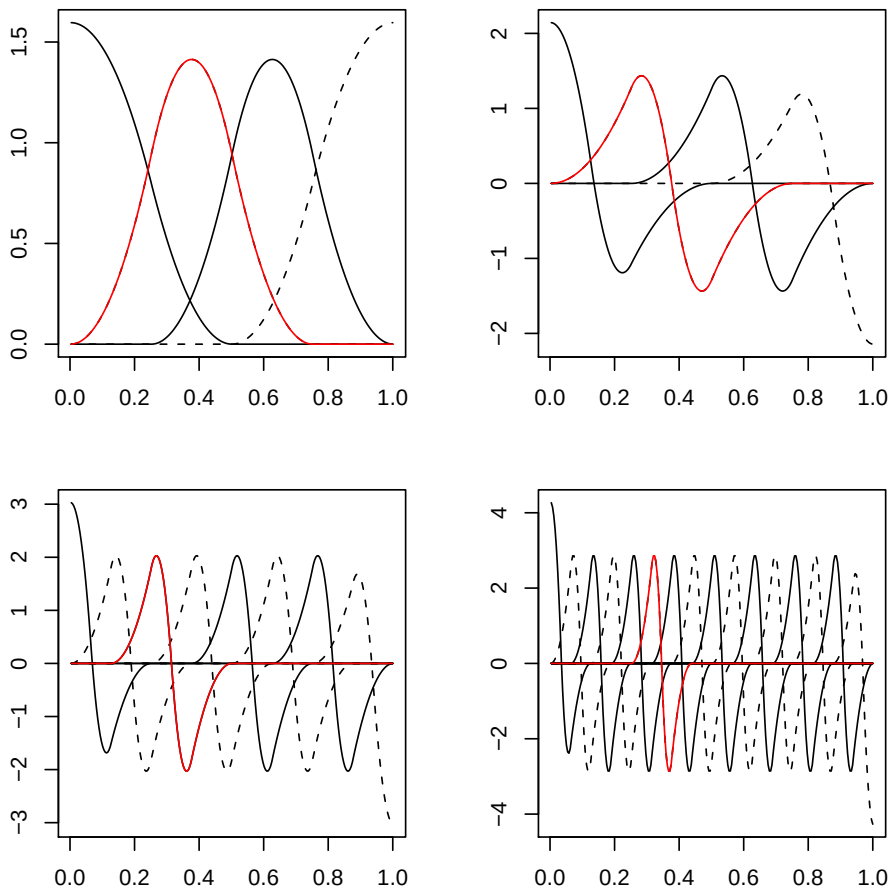


Multiresolution covariance model

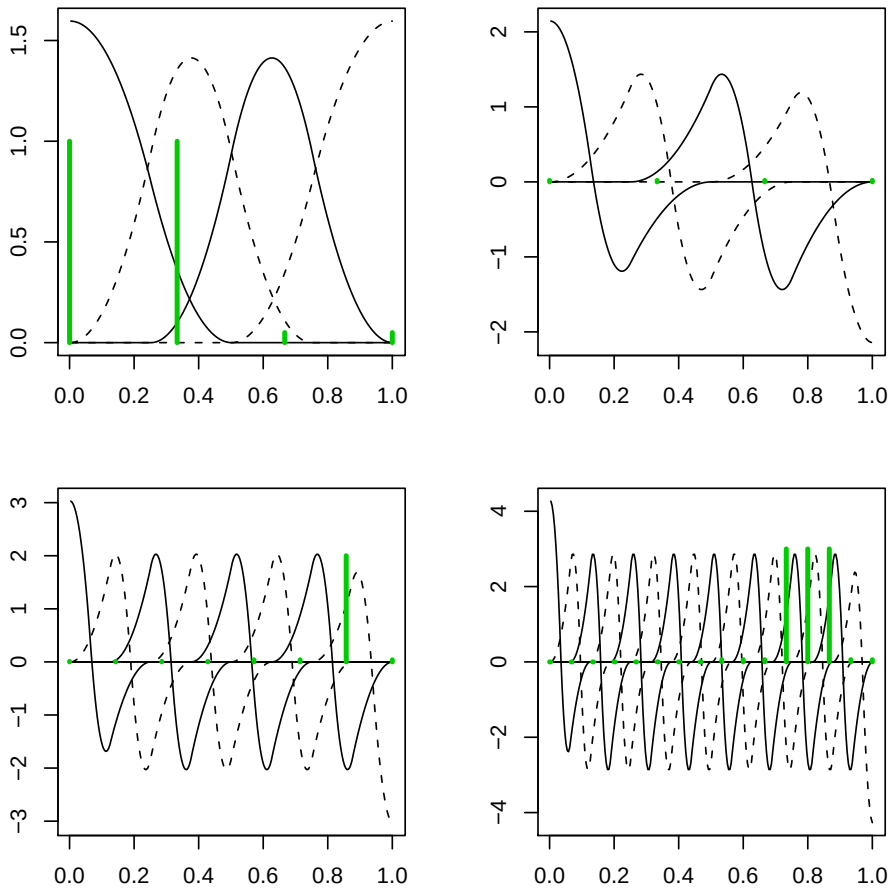
$$f(x) = \sum_{j=1}^{\infty} a_j \psi_j(x)$$

$\{\psi_j\}$ Wavelet basis function – fixed.
$\{a_j\}$ Gaussian with a small number of correlations – random.

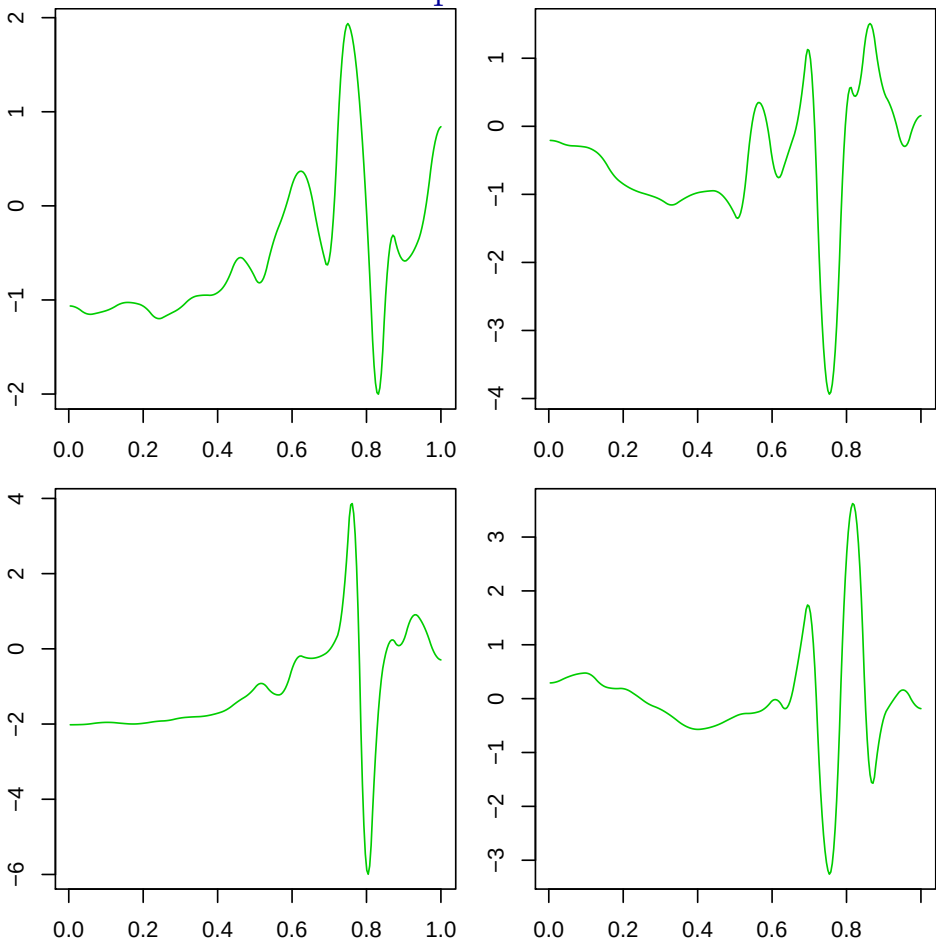
A 1-D wavelet basis



Some weights

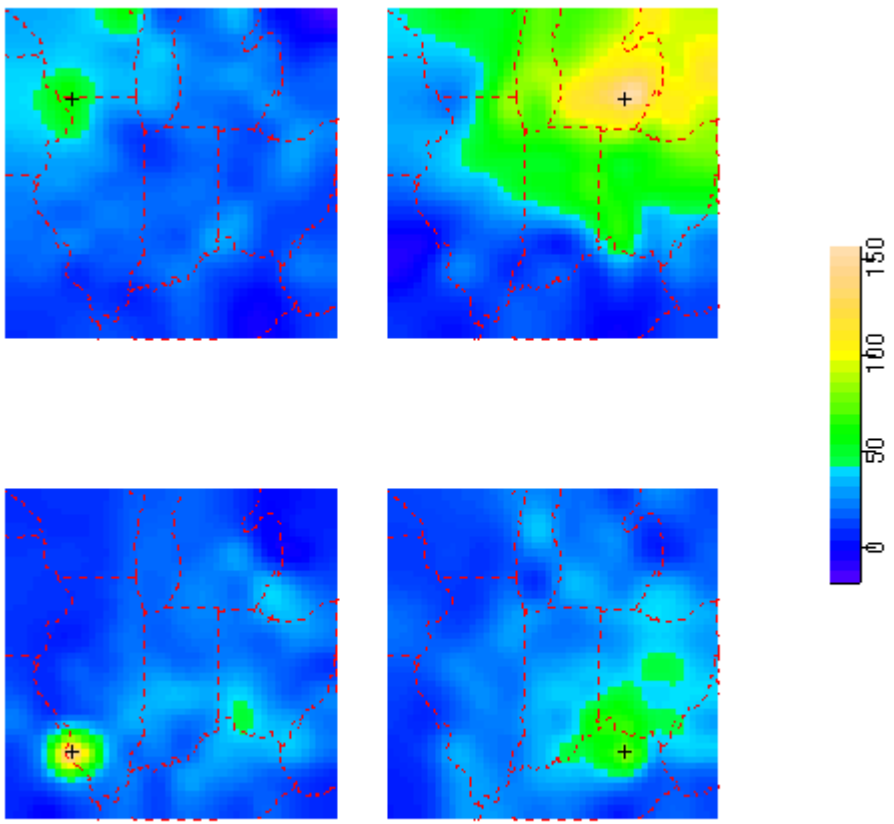


Four realizations of the process



Non stationary covariance for daily ozone

Estimated covariance function at four locations (approximately 90% decimation)



General Circulation Models (GCM)

Claudia Tebaldi, Judith Berner, Grant Brantstator

GCM

A deterministic numerical model that describes the circulation of the atmosphere. It is coupled to models for the ocean, ice , land etc. to simulate the entire climate system.

Different dynamic modes in a GCM

Scientific question: Does the atmosphere simulated by a GCM have different regimes i.e multiple equilibria?
If so, what are they and how long does the atmosphere spend in a particular regime?

Statistical problem: Given a multivariate, nonlinear time series

$$\mathbf{x}_t = A(\mathbf{x}_{t-1}) + \mathbf{e}_t$$

find A and partition the state space into regions of similar dynamics.

GCM Data

$\{\mathbf{x}_t\}$ is a 5 dimensional times series with 2M observations

- CCM0 state of the art GCM in early '80s

- No external forcing, "perpetual January", distinguishes between land and ocean, simple convection process.
- Resolution on a 7.5×4.4 degree grid (R15)
  model state vector is $\approx$ 18K, 30 minute time step
- 5-dimensional summary:
  coefficients of the first 5 EOFs for 300mb stream function
- Data series sampled twice per day over 1 million days

Neural net estimation

Recall: $\mathbf{x}_t = A(\mathbf{x}_{t-1}) + \mathbf{e}_t$

Estimate a nonlinear $\tilde{A}$ using the functional form

$$\tilde{A}(x) = \beta_0 + \sum_{l=1}^{h} \beta_l \Phi\left(\mu_l + \sum_{j=1}^{k} \gamma_{lj} x_{ij}\right)$$

where

$$\Phi(x) = \frac{e^x}{1 + e^x}$$

(single hidden layer)

"delinearizing" $A(x)$

Think of:

$$A(x_t) \equiv Bx_t + \tilde{A}(x_t)$$

where

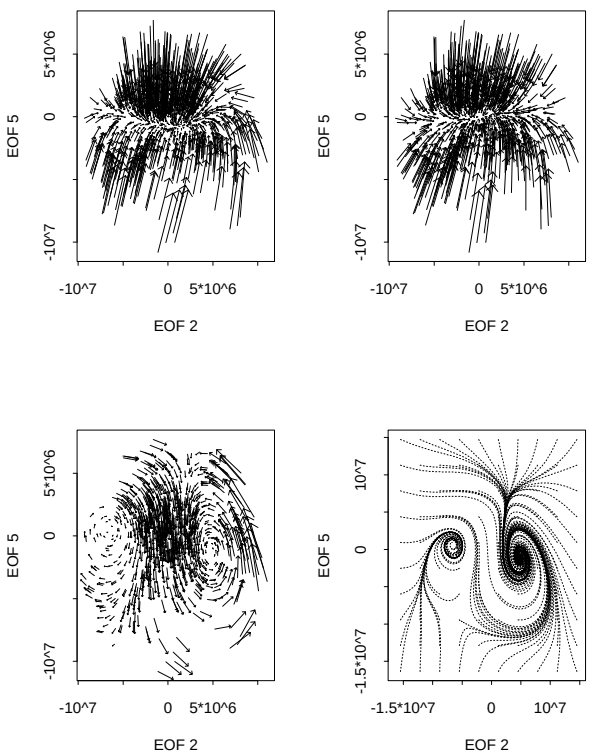
$$B = \left. \frac{\partial A(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=0}$$

and the difference from linearity:

$$\tilde{A}(\mathbf{x}_t) \equiv A(\mathbf{x}_t) - B\mathbf{x}_t$$

Looking at 2-d slices of the nonlinear component

Complete map, linear component
Nonlinear trajectories, Nonlinear motion

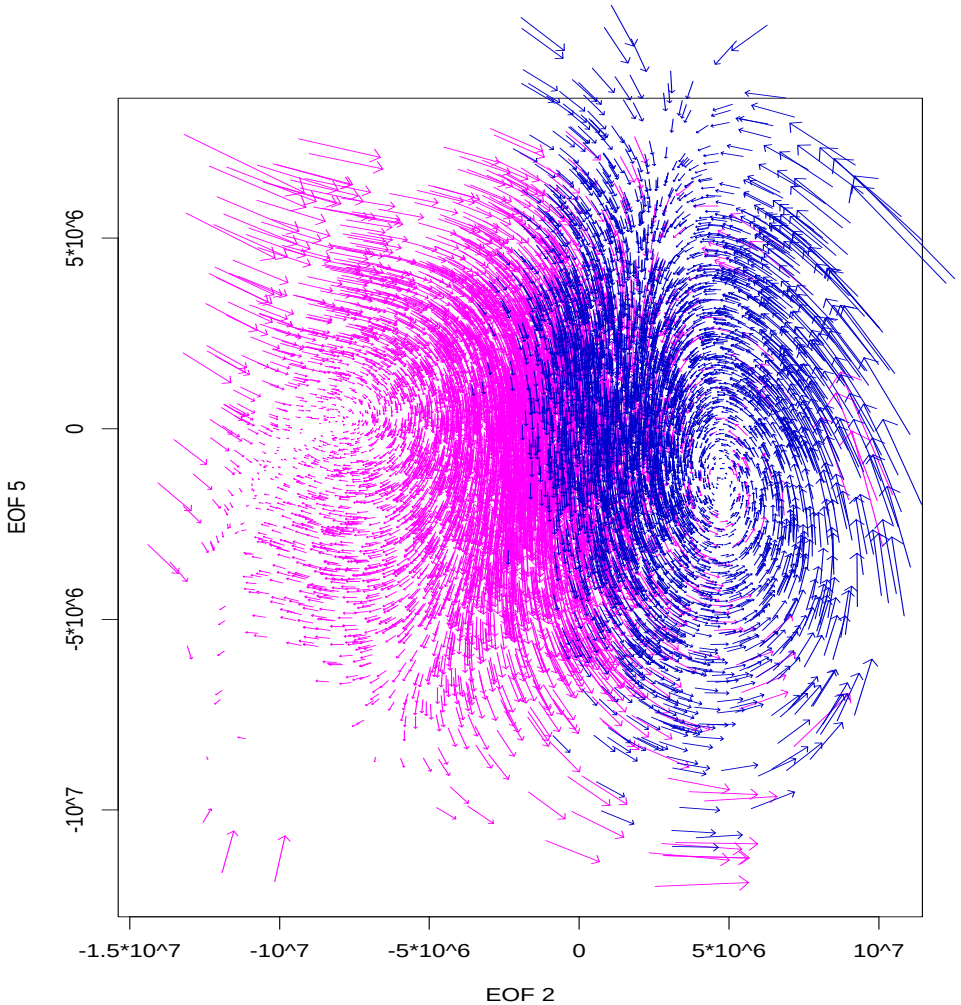


Finding Dynamical Clusters

- Compute the *Jacobians* of the nonlinear component at a *regular grid* of points in the phase space
- *Cluster* them
- use the resulting *clusters' centers* to cluster the *entire set of observations*

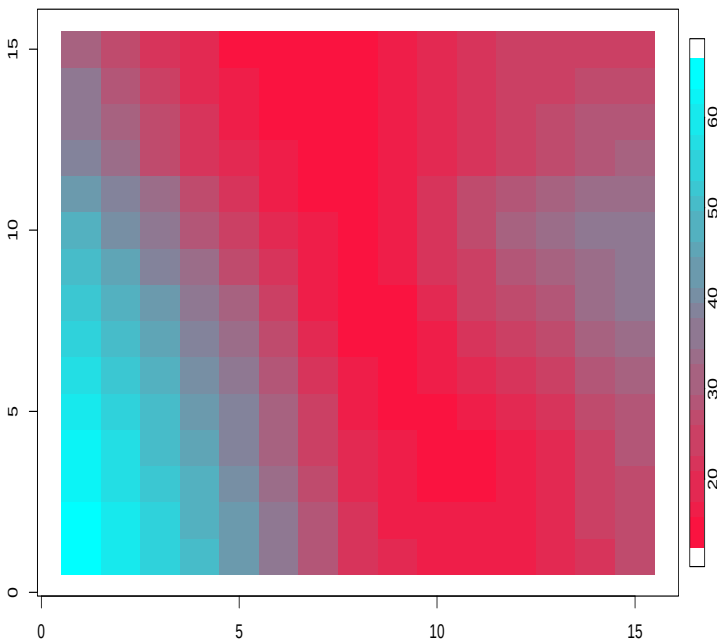
$$x_t \rightarrow \tilde{A}(x_t) \rightarrow \left[\frac{\partial \tilde{A}_i}{\partial x_j}\right]_{i=1,\ldots 5; j=1,\ldots 5} \rightarrow \text{clustering}$$

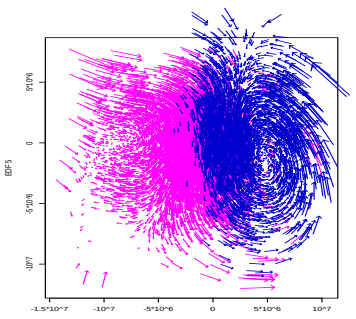
Agreement!



Cluster switching

Classify each point in the time series to a cluster. Plotted are expected times to switch for state vectors in the 2-5 EOF plane





Subgrid scale features

Enrica Bellone, Xaioqing Wu, William Collins, James Hack

Effects of clouds

- Recall GCMs are large scale models of the atmosphere. Sub-grid scale features, such as the overlap of clouds at different layers, are *parameterized* (their dynamics are not explicitly represented).
- Cloud resolving models (CRM) operate on scales of ≈ 1 km and resolve cloud-scale and mesoscale dynamics.
- Information from CRMs may be used to improve GCM parametrizations.

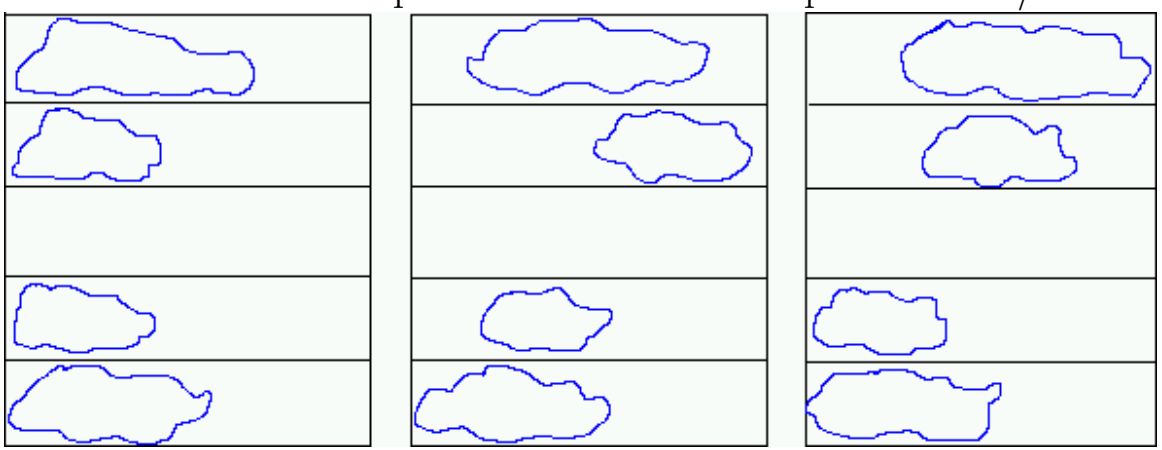
$$\text{CRM simulations} \xrightarrow{\text{statistical summaries}} \text{GCM parametrizations}$$

GCM Assumptions on Cloud Distribution

Cloud fraction for each vertical level is determined by the model. The vertical overlap between cloud layers must be specified.

Three common assumptions:

maximum overlap random overlap max/random overlap



- Within a cloudy region in a layer, the cloud amount is assumed homogeneous.

Markov Chain Approach

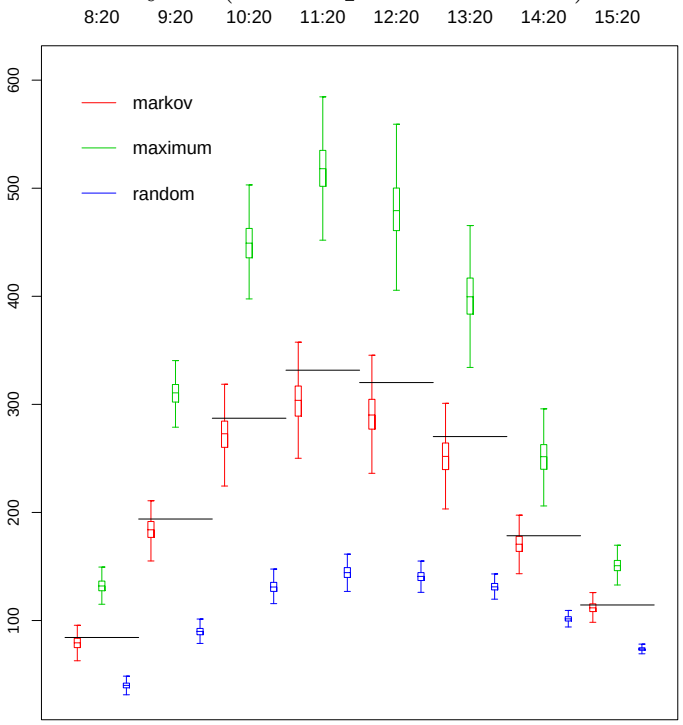
- Assume the (binary) cloud process follows a first order Markov chain in vertical height:
  $P(C_z^{(i,j)} = c_z \mid C_1^{(i,j)} = c_1, \ldots, C_{z-1}^{(i,j)} = c_{z-1}) =$
  $P(C_z^{(i,j)} = c_z \mid C_{z-1}^{(i,j)} = c_{z-1}) \equiv p_{c_{z-1}c_z z}$,
  where z is the height index and (i, j) are horizontal indices.
- The transition probabilities vary in height.
- No horizontal dependence is considered.

Radiative flux calculations

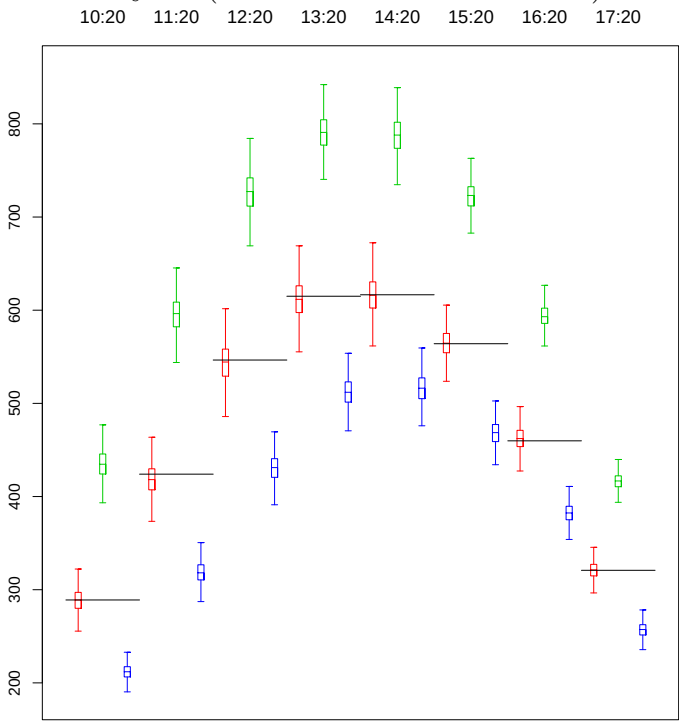
Generate cloud prescence/absence using three different overlap schemes and evaluate the radiation flux for a vertical column using the average cloud amount at each level based on the CRM simulation.

Downward Shortwave Flux at the Surface

Day 2 (non-squall clusters)



Day 7 (scattered convection)



Units: W/m^2.

Summary

- There are different uses of statistics with numerical output:
  Simplified prediction models, Detection of regime changes, Stochastic summaries of complex phenomenon
- The tools are varied:
  Wavelet/Multiresolution bases, Nonparametric regression and clustering, Markov models.
- Nesting each of these examples in a Bayesian model will produce companion measures of uncertainty.
- Workshop on *Space-time Statistics*, June 1-6, Boulder, CO