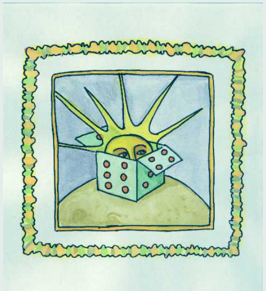
Data Mining in the Atmospheric Sciences

Douglas Nychka, Sarah Streett and Linda Mearns
Geophysical Statistics Project,
National Center for Atmospheric Research
`http://www.cgd.ucar.edu/stats/`

- National Assessment and Climate Change
- Weather generators and hierarchical models
- General Circulation Models
- Regimes in the atmosphere

Data mining: Discerning empirical, possibly complex relationships from large data sets.



National Center for Atmospheric Research

≈ 1000 people total, several hundred PH D (physical) scientists,
half the budget (≈ 60M) is a single grant from NSF-ATM

Research on nearly every aspect related to the atmosphere

Climate, Weather, the Sun, Ocean/atmosphere, Ecosystems, Economic impacts, Air quality, Instrumentation, Scientific computing and ...

Statistical methods for the geosciences

Motivation

Climate is the average of ”weather” over long time scales.

$$\text{Climate}(\mathbf{x}) = E[\text{weather}(\mathbf{x})]$$

Premise: Global warming (climate change) is occurring ... and most scientists attribute some of the warming to increasing levels of greenhouse gases.

Problem: Translate geophysical predictions of climate change into terms of daily weather. How do (sometimes subtle) changes in weather effect society, the economy and the environment.

Strategy: Build weather generators from observational data and climate change scenarios. Feed generated weather to numerical, *impact* models to assess the effects of a changing climate.

In this work: *Corn yields for the Southeast US.*

Schematic of approach

Current Climate → WGEN → Crop model yields

Modified Climate → Modified WGEN → Crop model yields

yield differences

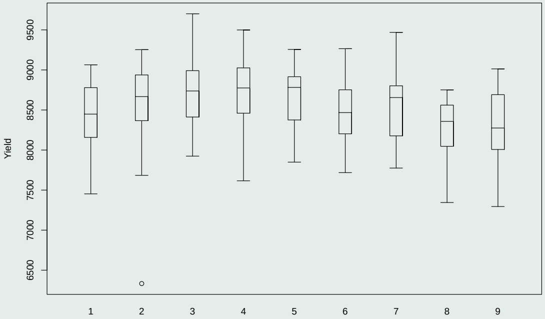
The observational weather record

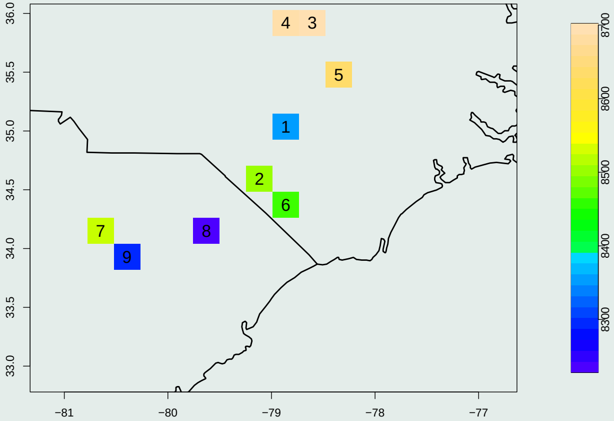
A large data set ($> 15M = 3 \times 30 \times 365 \times 500$) , with

- irregular spatial locations (for SE US $\approx$ 500 stations)
- irregular observation periods
- weirdness: outliers, min temp $>$ max temp

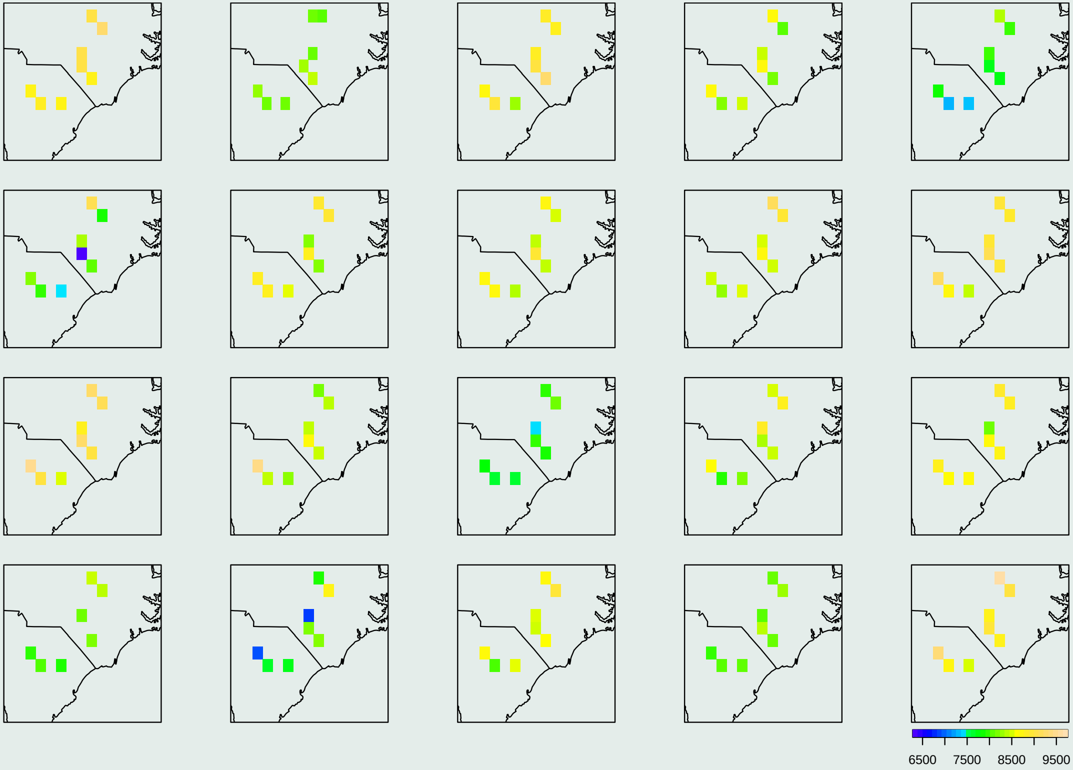
CERES corn model

Average yield (Kg/Ha) using observed daily weather 1965-1984





Spatial sequence of yields: CERES corn model (1960-1984)



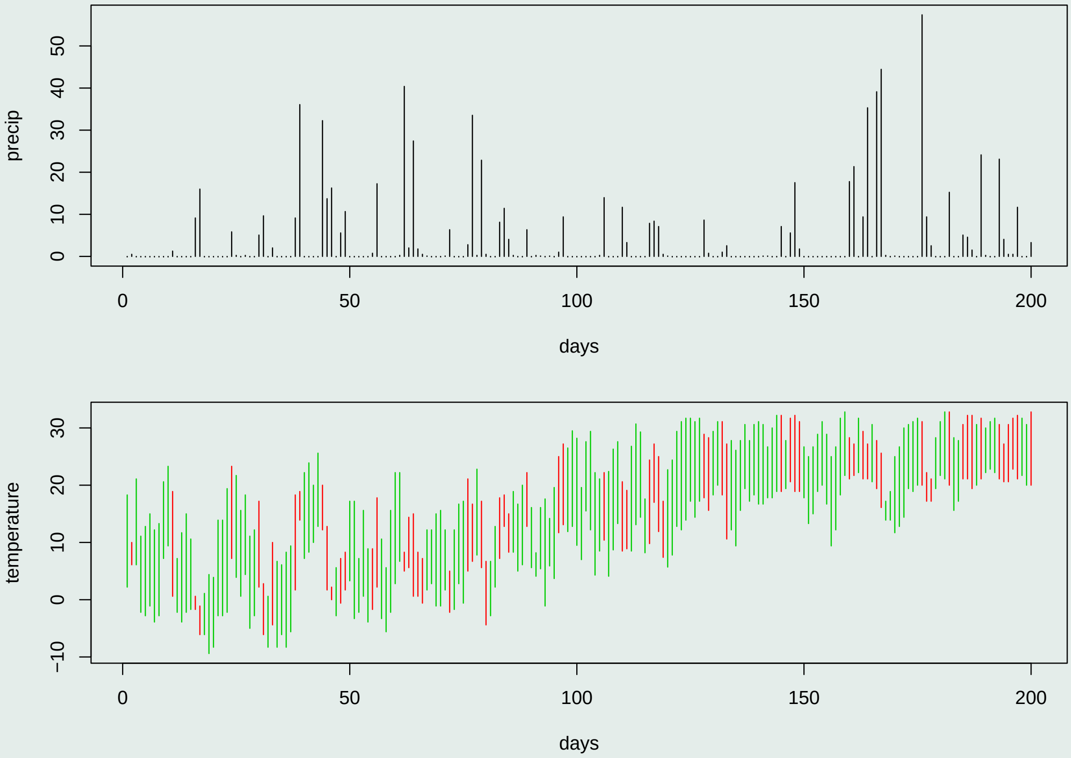
Weather generators

WGEN Richardson (1981), Parlange and Katz (1999)
Set $\mathbf{Z}_t$ to be the daily weather variables. Essentially a multivariate time series model for

- Precipitation: occurrence, amount
- Solar Radiation
- Daily average temperature and range
- Humidity and Wind speed

The key is to *separate the model into dry and wet days*.

Weather for first 200 days for station 1 conditioned by occurrence



A model for precipitation occurrence

Occurrence $Y_t =$ (0 or 1) follows an observation driven model:

$$P(Y_t = 1) = p_t$$

where p_t depends on past values of Y and seasonality.

Let U, be a uniform R.V. on $[0, 1]$:
if $U > p_t$ *no rain,* if $U \leq p_t$ *rain*

Modeling Hierarchy:

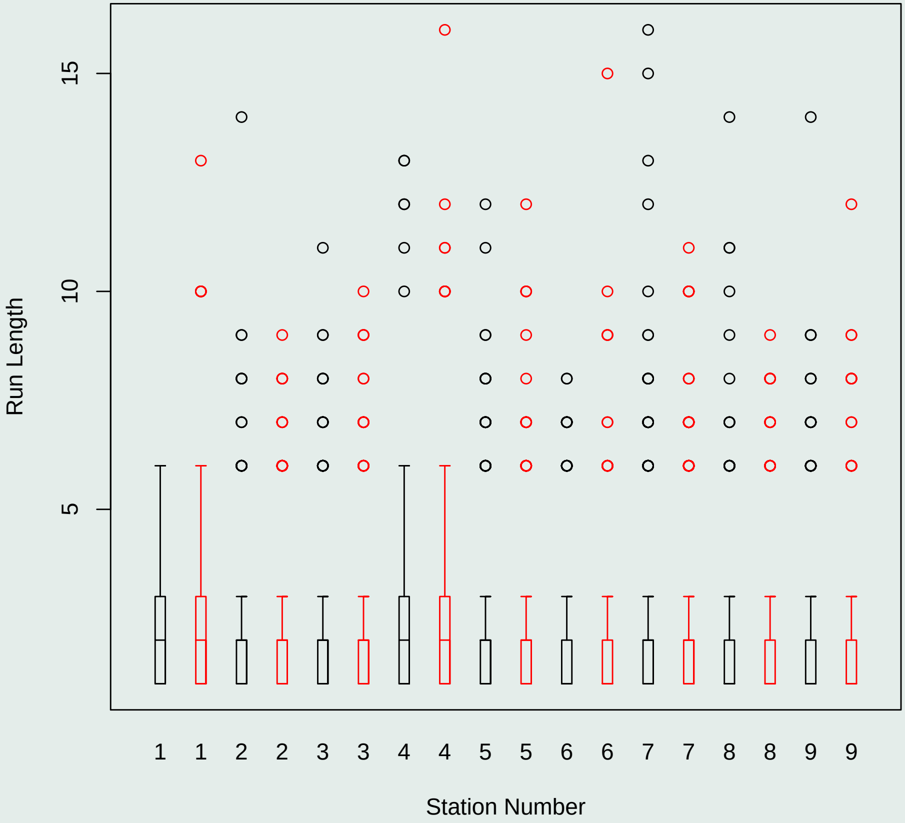
logit transformation $p_t = e^{\theta_t}/(1 + e^{\theta_t})$

seasonality and memory $\theta_t = \mathbf{x}_t\beta + \epsilon_t$

means depend on past $\epsilon_t = \alpha(Y_{t-1}, Y_{t-2}) + \delta U_{t-1}$

innovations depend on past $U_{t-1} = (Y_{t-1} - p_{t-1})/\sqrt{p_{t-1}(1 - p_{t-1})}$

Evaluate the occurrence model by checking the distribution of “wet spells” against the observed station data. (black= model red= data)



Generating Precip Amount, Solar radiation, temperature, etc.

Given that it has rained, the rain amounts are assumed to follow a Gamma distribution where the gamma parameters vary over season.

Condition on occurrence, find (seasonal) transformations of the variables to standard normals. $\mathbf{u}_t = \Gamma_t(\mathbf{Z}_t)$.
Γ based on best fitting Gamma distribution followed by a (nonparametric) spline transformation.

$\mathbf{u}_t$ evolves according to a (seasonal) AR 1.

$$\mathbf{u}_t = A_t \mathbf{u}_{t-1} + \mathbf{e}_t$$

Adding spatial structure

Spatial dependence

How does one add stochastic structure that is coherent over space?

Precipitation occurrence (0,1) process:

$$P(Y_t(\mathbf{x}) = 1) = P(U_t(\mathbf{x}) < p_t) = P(\Omega_t(\mathbf{x}) < F^{-1}(p_t))$$

with $U_t(\mathbf{x})$ a correlated spatial process with marginals that are uniform.

We assume that $\Omega_t(\mathbf{x}) = F^{-1}(U_t(\mathbf{x}))$ a Gaussian spatial process $F \sim \mathrm{N}(0, 1)$.

AR 1 innovations:
Assume that $\mathbf{e}_t(\mathbf{x})$ is a multivariate Gaussian spatial process.

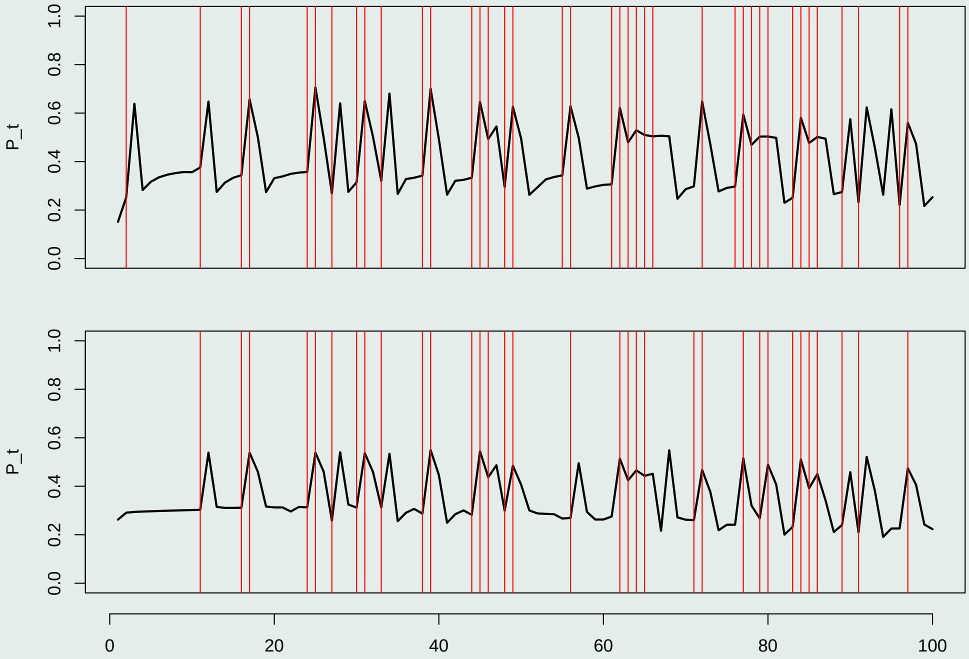
Extrapolating/Smoothing WGEN parameters

Smooth or interpolate weather generator parameters over space, These include:

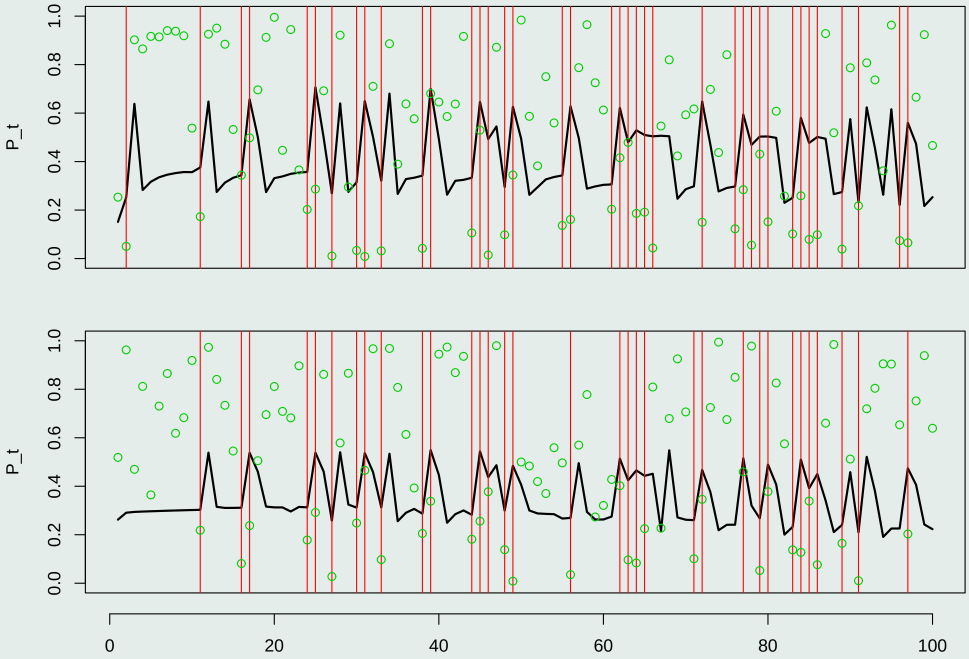
- Seasonality in means and standard deviations
- Transformations to Normality
- Autocorrelations
- Parameters of Precipitation occurrence models

Use functional data methods and spatial statistics (???).

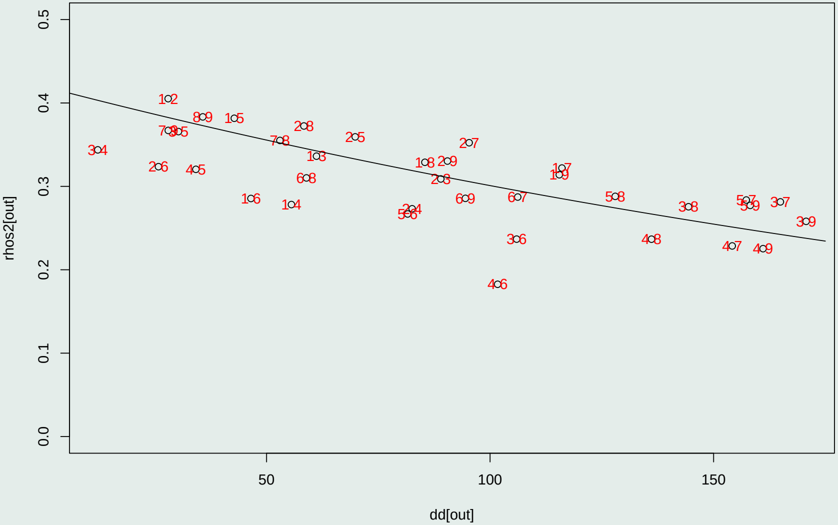
Observed occurrences and p_t for stations 1 and 2



Observed occurrences and p_t for stations 1 and 2



Correlations among all stations against distance



Correlation model: $COR(\mathbf{x}, \mathbf{x}') = .42e^{-(\|\mathbf{x}-\mathbf{x}'\|/300)}$

Summary

- The seemly straightforward exercise of building a weather generator poses new statistical models.
  e.g. *functional data*, *NonGaussian*, *Space-time Processes*
- Linking models over a region involves mining relationships among WGEN's for individual locations.

Finding different regimes in a climate model.

Many numerical models of geophysical systems provide very large and complex data sets. The the model output has rich structure that is often not examined in detail by modelers. It is an important opportunity for data mining tools to identify complicated model properties that are not obvious through simple summary statistics.

General Circulation Models (GCM)

- *GCM*: A deterministic numerical model that describes the circulation of the atmosphere. It is coupled to models for the ocean, ice , land etc. to model the entire climate system.
- Conceptually based on grid boxes (for the NCAR climate system model: there are $128 \times 128 \times 17$) and the state of the atmosphere is the average quantities for each box .

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. Model output from a run of CCM0

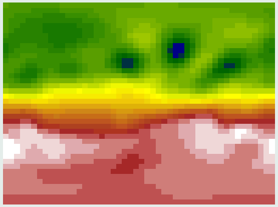
- CCM0 state of the art GCM in early ’80s

- 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

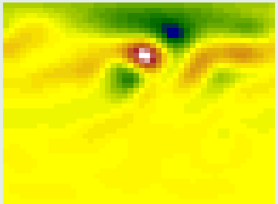
300mb stream function is the nondivergent part of the horizontal wind field at 300mb. (The Laplacian of the stream function is the vorticity. It is a useful summary of the flow in the Northern Hemisphere mid-latitudes

The five EOFs used to reduce the data

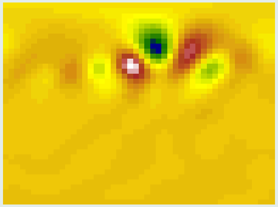
EOF1



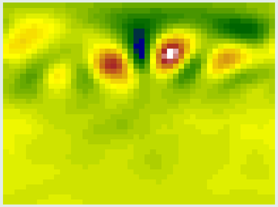
EOF2



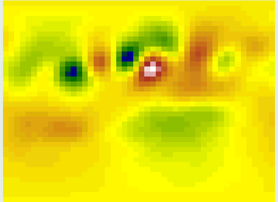
EOF3



EOF4



EOF5



Fitting an multivariate AR model to the state vector

$A(\mathbf{x})$ is overwhelmingly linear ...
but with $2 * 10^6$ data points we can be picky!

Neural net estimation

Estimate a nonlinear A using the functional form

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

where

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

(single hidden layer)
This is not quick ...
Training set: 600K observations
Computer time ≈ 6 hours on a shared SGI Origin2000

“DELINEARISING” $A(x)$

Think of:

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

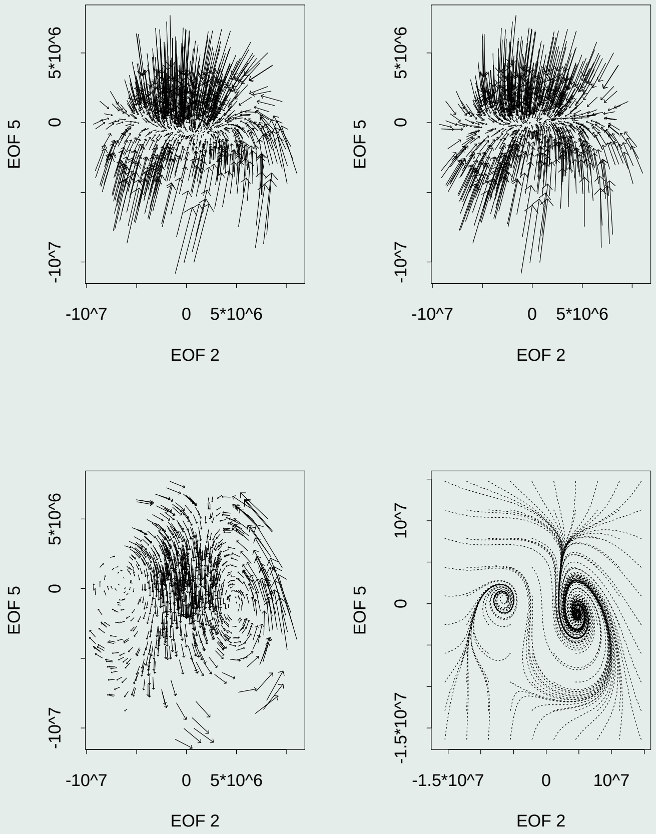
where

$$B = J_A(0)$$

is the Jacobian of the map $A(.)$ computed at 0 then:

$$\tilde{A}(\mathbf{x}_t) \equiv A(\mathbf{x}_t) - J_A(0)\mathbf{x}_t$$

Looking at 2-d slices of the nonlinear map

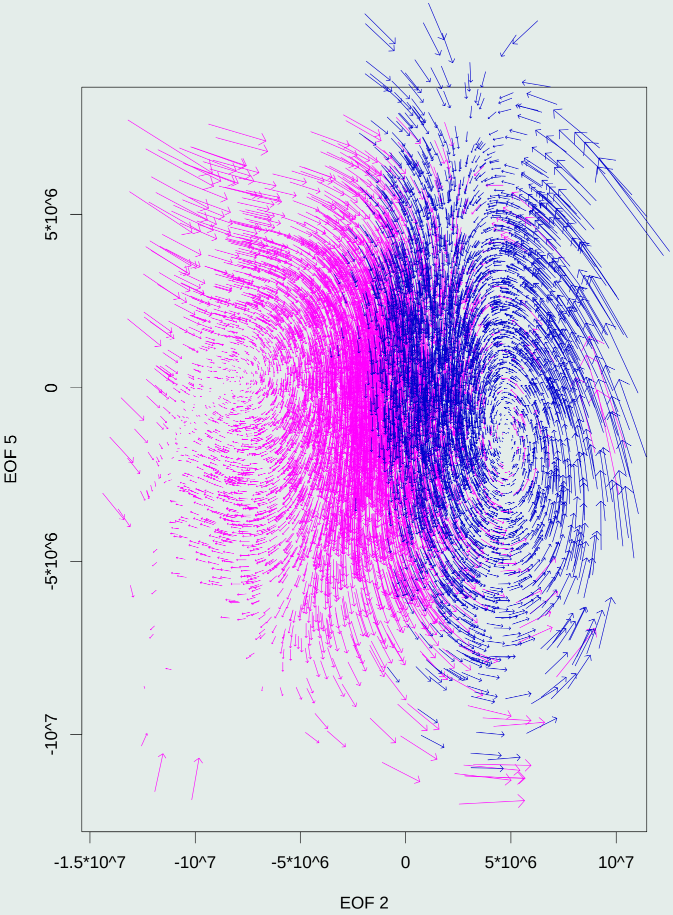


FINDING DYNAMICAL CLUSTERS

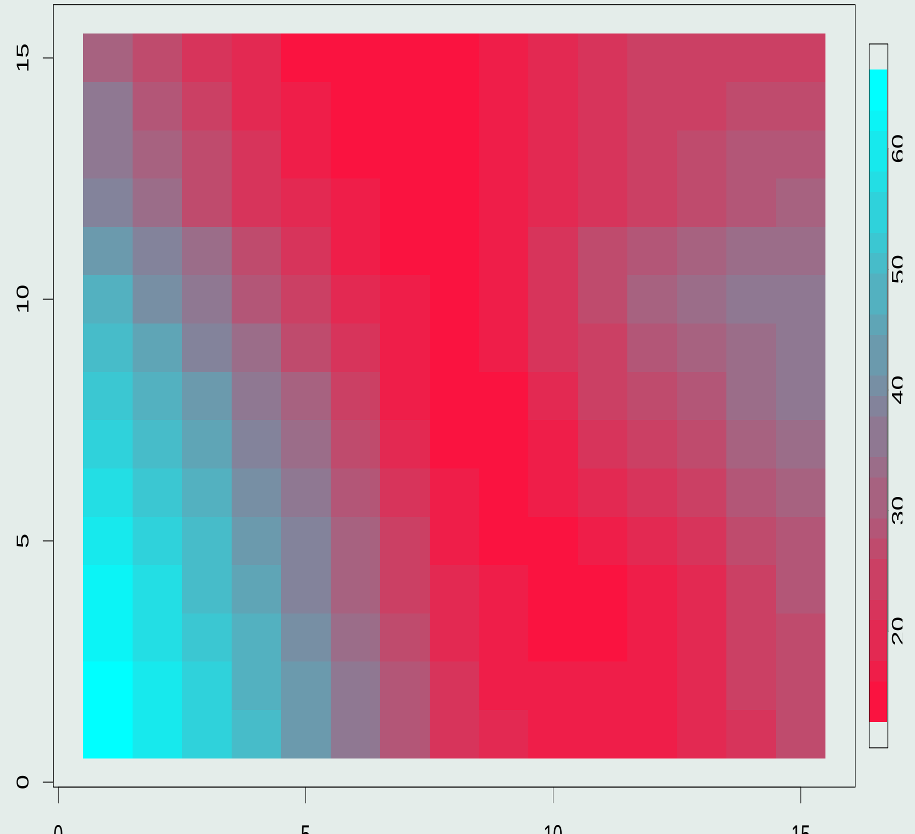
- Compute the *Jacobians*
  of the nonlinear component
  at a *regular grid* of points in the phase space
- *Cluster* them
- use the resulting *clusters' centers* to cluster the *entire set of observations*

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

CLUSTERS AND SWIRLS AGREE!



Classify each point in the time series and estimate the expected times of switch from one cluster to another. Plotted are times to switch for state vectors in the 2-5 EOF plane



Summary

- We have *estimated and quantified the nonlinear component* of the map

$$X_{t-1} \rightarrow X_t;$$

- Dynamics of CCM0 can be productively clustered into at least two distinct regimes.