Finding different regimes in a climate model.

Douglas Nychka
Geophysical Statistics Project
National Center for Atmospheric Research (NCAR)
`www.cgd.car.edu/stat` Claudia Tebaldi, Athene Software
Grant Brantstator, NCAR
Judith Berner, NCAR

Outline

- Climate Models
- Dynamical systems
- Neural networks
- Clustering

Overview

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 .
- Each grid cell is $\approx$ 200km $\times$ 200km in area and important processes that affect large scale flow are not resolved by the grid.

Modeling the atmosphere

The physical equations to describe atmospheric motion are derived from fluid mechanics and thermodynamics.

The complete state depends on:

- 3-d wind field, $\boldsymbol{v}$
- pressure p
- temperature T
- heating by radiation $Q_{\mathrm{r}ad}$, condensation $Q_{\mathrm{C}on}$
- evaporation E and condensation C from clouds
- D_H D_M and D_q are diffusion terms.

Primitive Equations

$$
\begin{aligned}
\frac{\partial \boldsymbol{v}}{\partial t} &= -\boldsymbol{v} \cdot \nabla \boldsymbol{v} - \omega \frac{\partial \boldsymbol{v}}{\partial p} + f \boldsymbol{k} \times \boldsymbol{v} - \nabla \Phi + \boldsymbol{D}_M \\
\frac{\partial T}{\partial t} &= -\boldsymbol{v} \cdot \nabla T + \omega \left(\frac{\kappa T}{p} - \frac{\partial T}{\partial p} \right) + \frac{Q_{\mathrm{r}ad}}{c_p} + \frac{Q_{\mathrm{C}on}}{c_p} + D_H \\
\frac{\partial q}{\partial t} &= -\boldsymbol{v} \cdot \nabla q - \omega \frac{\partial q}{\partial p} + E - C + D_q \\
\frac{\partial \omega}{\partial p} &= -\nabla \cdot \boldsymbol{v} \\
\frac{\partial \Phi}{\partial p} &= -\frac{RT}{p}
\end{aligned}
$$

where Φ geopotential (height) and $\omega = \frac{dp}{dt}$.
Despite the precision of this physical description, the radiative and cloud terms involve complicated additional physics.

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

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

estimate A and use it to partition the state space into regions of similar dynamics.

Here I showed an mpeg animation of the Lorenz system to indicate different modes of the system: butterfly wings verses the body. See `www.cgd.ucar.edu/stats/data/lorenz` for this animation.

GCM Data

$\{\boldsymbol{x}_t\}$ 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

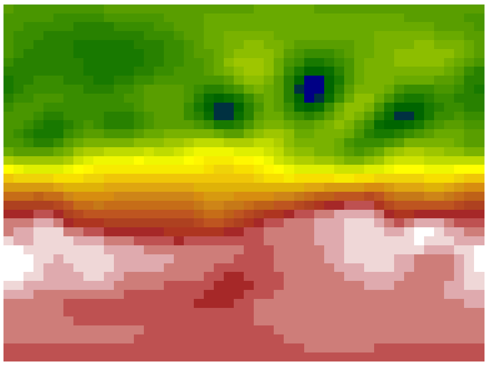
- No external forcing, "perpetual January", distinguishes between land and ocean, simple convection process.
- Resolution on a 7.5×4.4 degree grid (R15)
  model state vector is $\approx$ 18K, 30 minute time step
- 5-dimensional summary:
  coefficients of the first 5 EOFs for 300mb stream function
- Data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

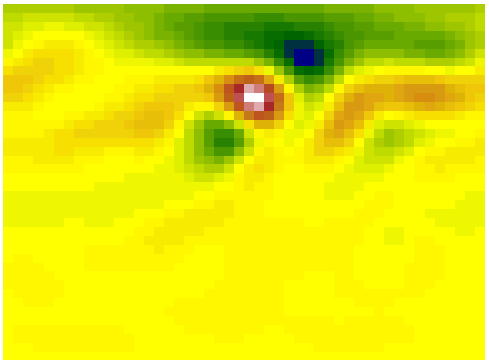
Mpeg animation of this field over time for 500 days. See `www.cgd.ucar.edu/stats/data/ccm0` for this animation

WHAT DO THE EOFs LOOK LIKE?

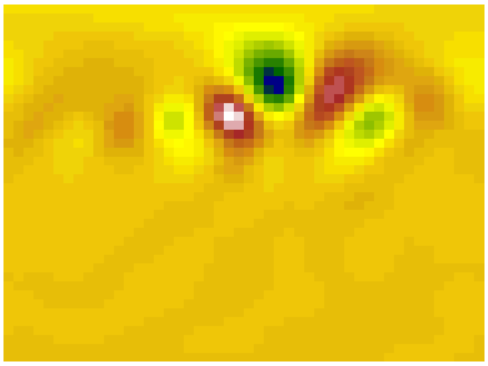
EOF1



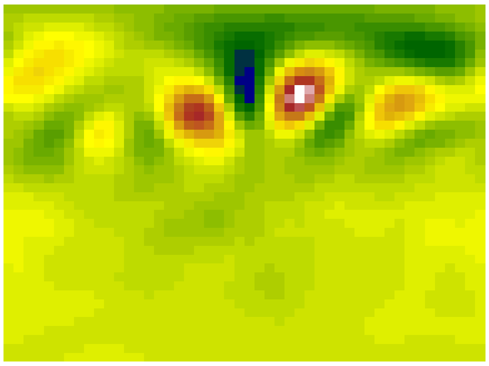
EOF2



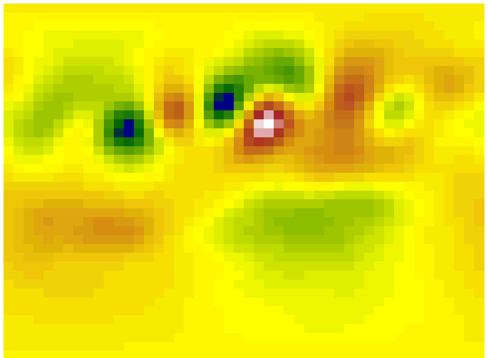
EOF3



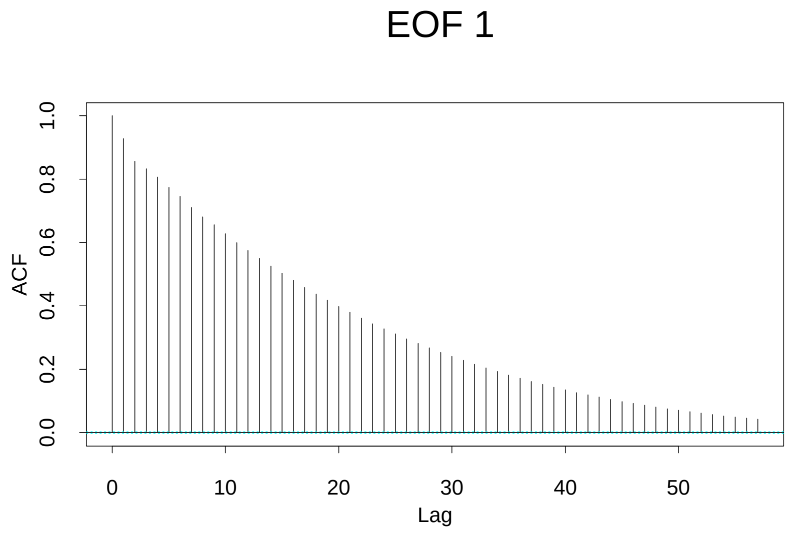
EOF4

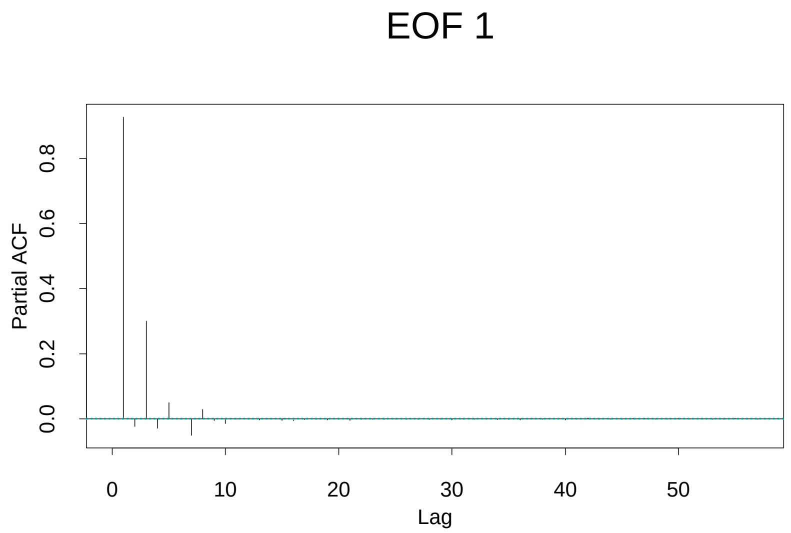


EOF5



$A(\boldsymbol{x})$ IS OVERWHELMINGLY LINEAR



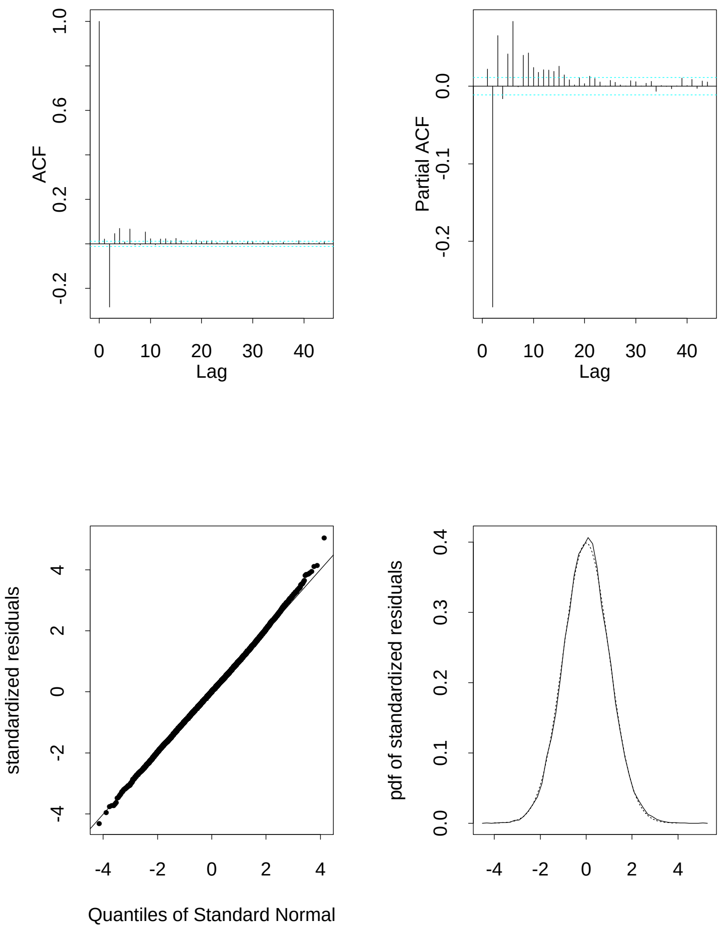


$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \epsilon_t$$

$$\epsilon_t \sim N(0, \sigma^2 I)$$

$$R^2 = .88$$

WHAT IS LEFT:



WITH $2 * 10^6$ DATAPOINTS WE CAN BE PICKY!

NEURAL NETWORKS 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

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

This is not quick ...
Training set: 600K observations
Computer time $\approx$ 6 hours on a shared Origin2000

“DELINEARISING” $A(x)$ Think of:

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

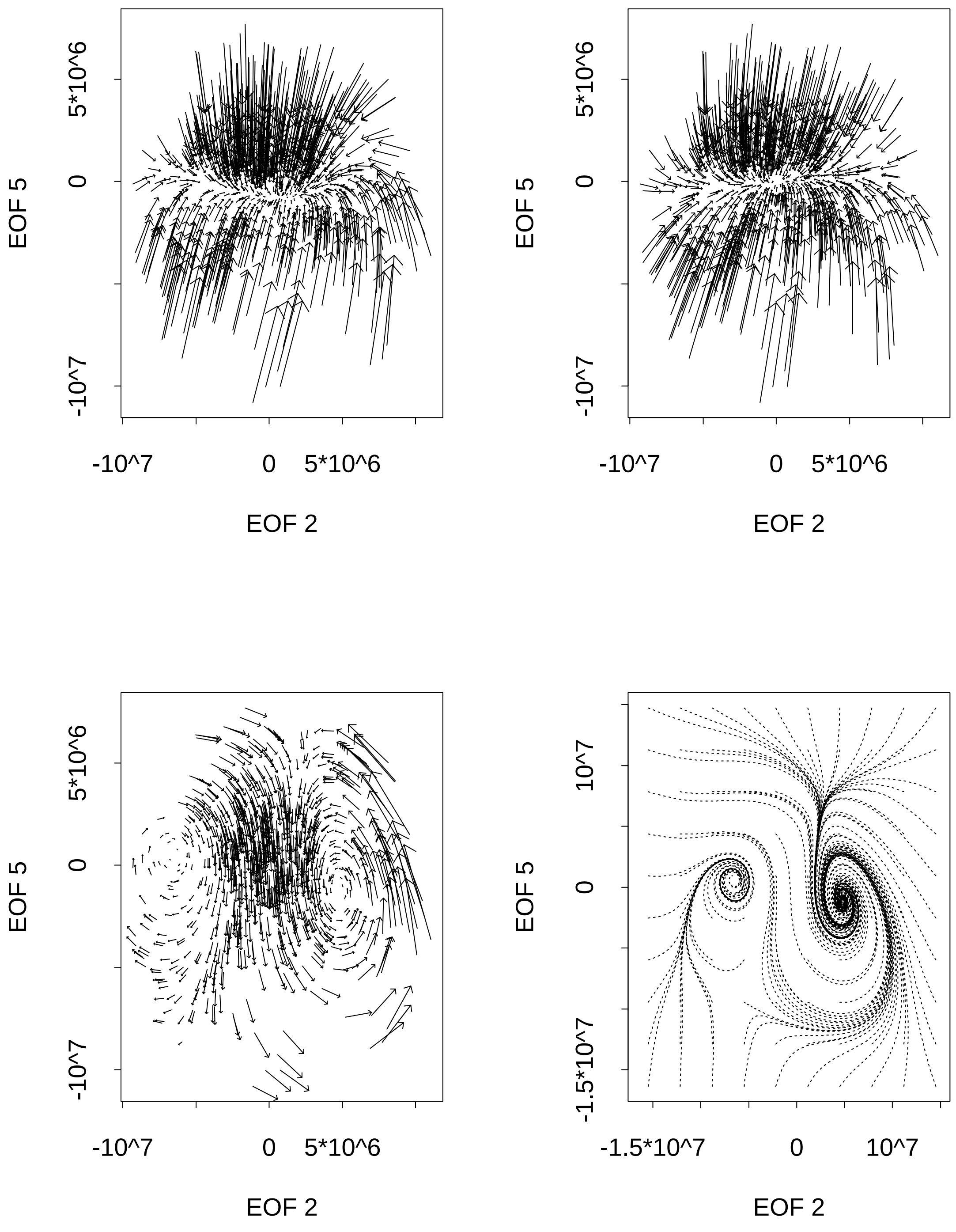
where

$$B = J_A(0)$$

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

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

Looking at 2-d slices of the nonlinear map

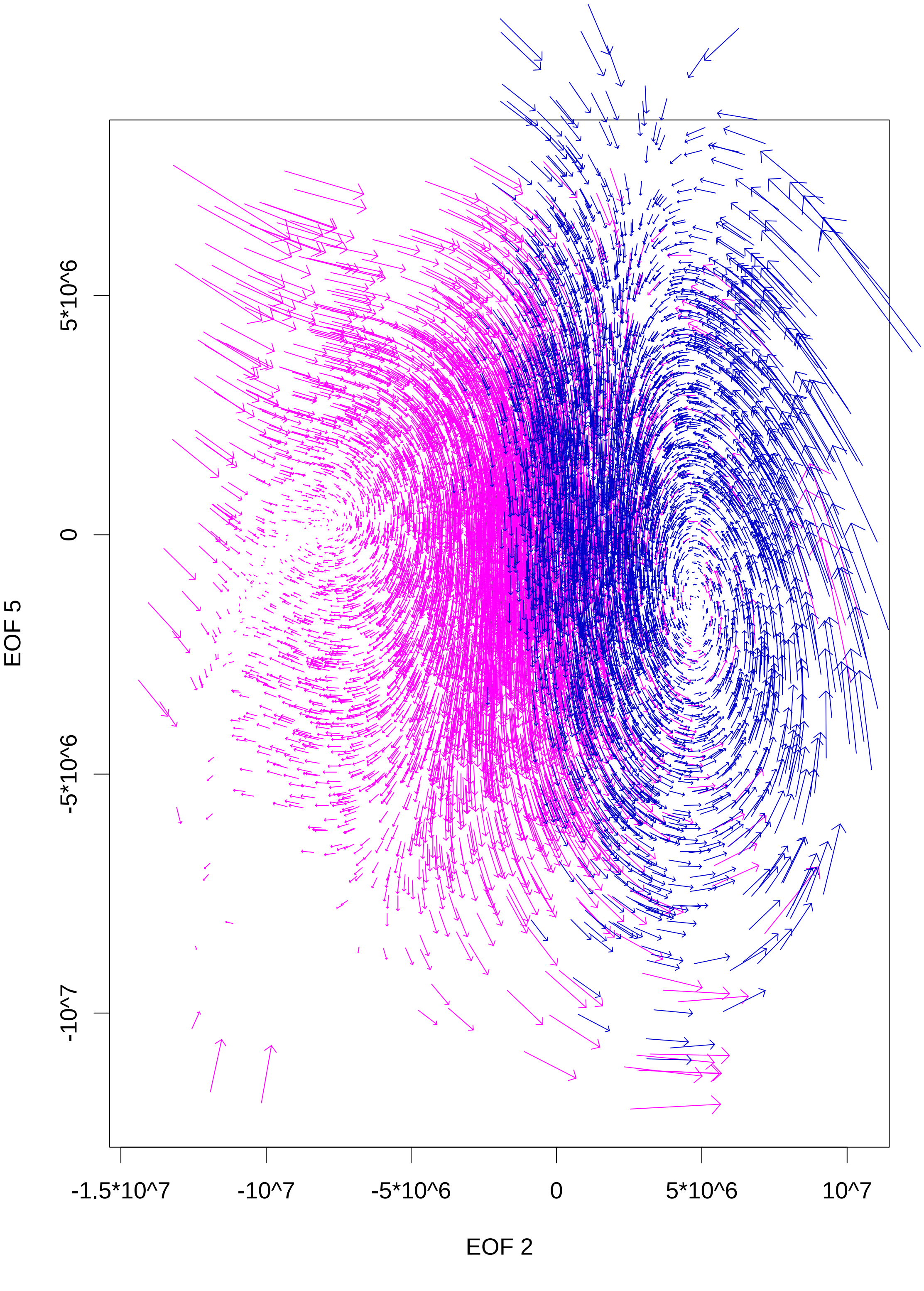


FINDING DYNAMICAL CLUSTERS

- 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*

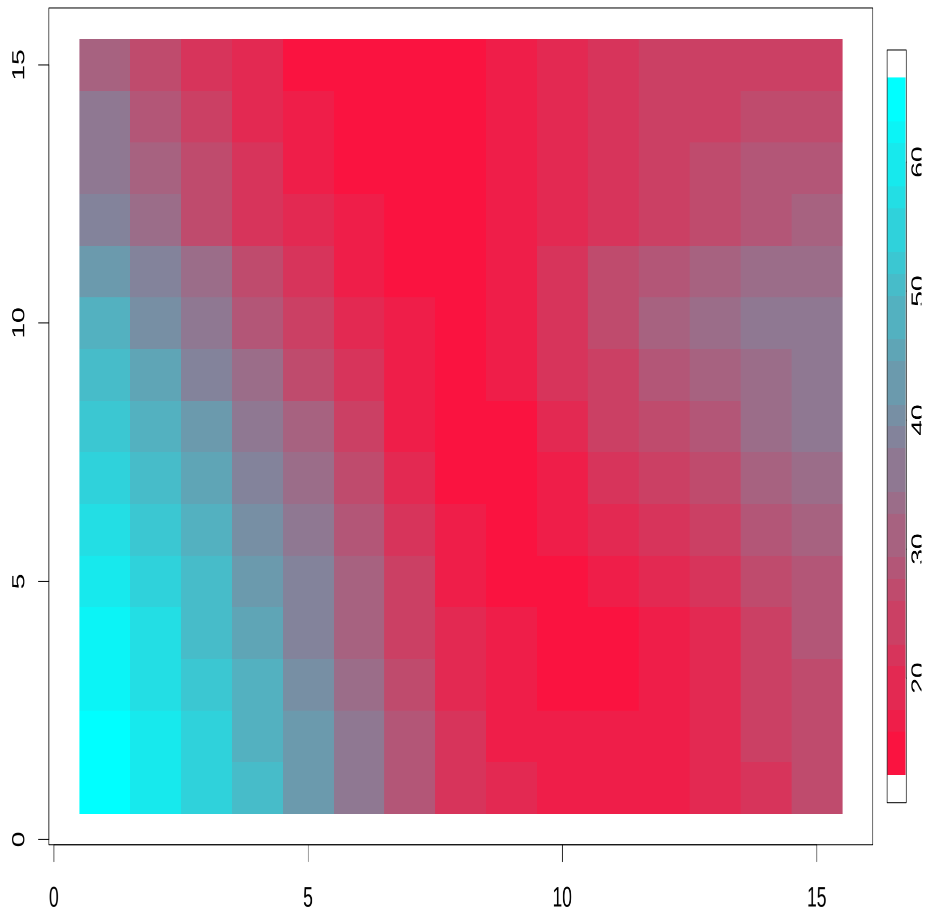
$$x_t \to \tilde{A}(x_t) \to \left[\frac{\partial \tilde{A}_i}{\partial x_j}\right]_{i=1,\ldots 5; j=1,\ldots 5} \to \text{clustering}$$

CLUSTERS AND SWIRLS AGREE!



SOME ANALYSIS OF THE CLUSTERS

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



CONCLUSIONS

- We have *estimated and quantified the nonlinear component* of the map
$$X_{t-1} \to X_t;$$
- We have *subdivided the phase-space* with respect to it;
- We have *estimated boundaries*, computed *expected time of switch*, approximated the dynamics by *piecewise linear models*;
- All amounts to a .4% increase in the explanatory power of a simple linear map. Is it significant?
- It is *statistically*
- It is *scientifically*.