

# Long-Range Climate Forecasts Using Data Clustering and Information Theory

Dimitris Giannakis

Courant Inst. of Math. Sciences, New York University  
dimitris@cims.nyu.edu

NY Workshop on Computer, Earth, and Space Sciences  
Goddard Institute for Space Studies  
February 25, 2011

# Acknowledgments

Joint work with Andrew Majda (Courant)

Discussions

Boris Gershgorin (Courant)

Illia Horenko (Univ. Lugano)

Dataset

Rafail Abramov (Univ. Illinois Chicago)

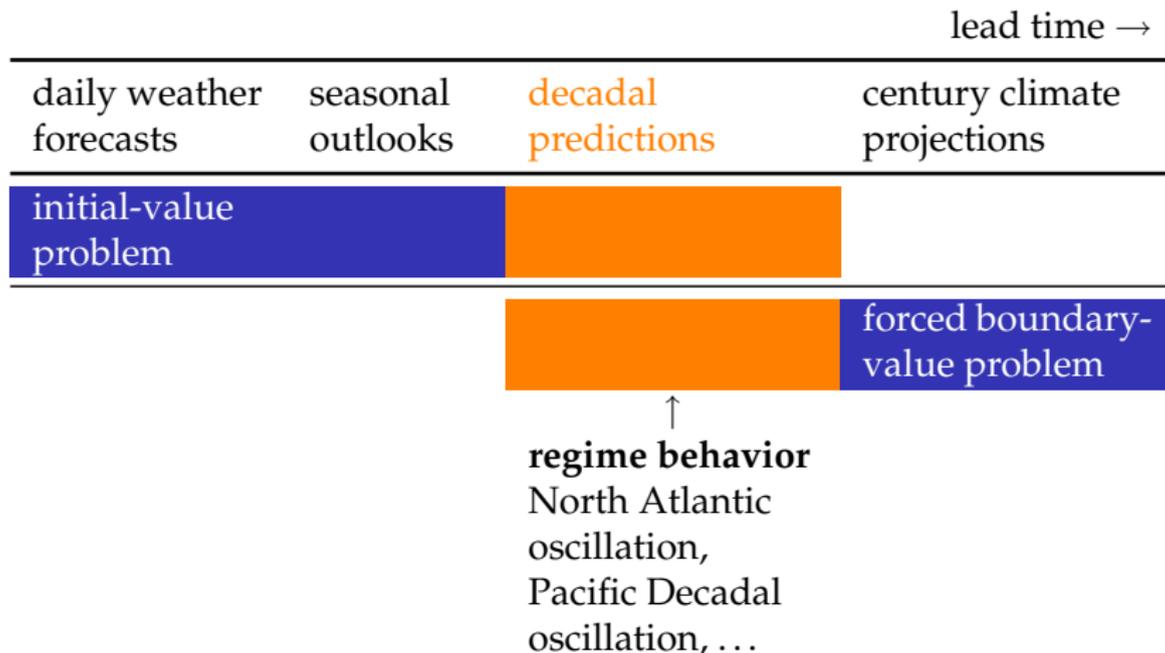
Computing resources

Paul Fischer (Argonne Natl. Lab)

# Outline

- 1 What is predictability?
  - Initial-value vs. forced-response problem
  - Predictability as information gain beyond equilibrium climate
- 2 Revealing long-range predictability via information theory and data clustering
  - Strategies for partitioning the set of initial data
  - The information content in coarse-grained initial data
- 3 Long-range forecasts in a double-gyre ocean model
  - Predictability of large-scale observables beyond their decorrelation time
  - Model error in Markov models of regime transitions

# Initial-value vs. forced-boundary forecasts



After Meehl *et al.* (2009), *Bull. Amer. Met. Soc.*, **89**, 303.

# Pacific decadal oscillation (PDO)

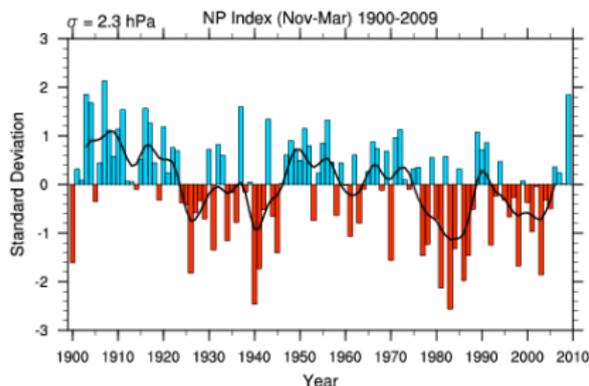
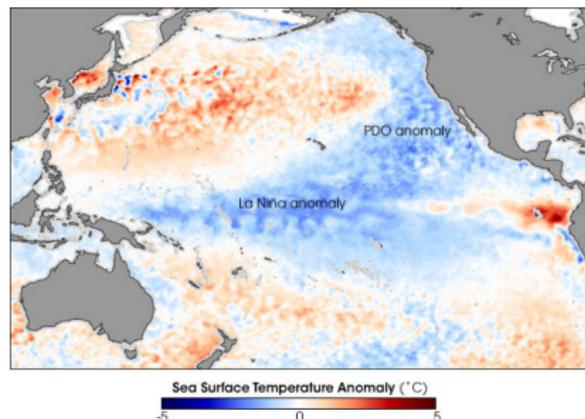


Image sources: [earthobservatory.nasa.gov/IOTD](http://earthobservatory.nasa.gov/IOTD),

[www.cgd.ucar.edu/cas/jhurrell/indices.info.html](http://www.cgd.ucar.edu/cas/jhurrell/indices.info.html)

- A long-lived *El Niño*-like pattern classified as being in either **warm** or **cool** phases.
- Cool phase (shown left) is associated with above-average precipitation in the Eastern US.
- Only 2–3 cycles in the last century.
- Dynamical origins and predictability are an area of active research.

# Setting

- High-dimensional, chaotic, and strongly mixing dynamical system

$$\frac{dx}{dt} = F(x, t), \quad x \in \mathbb{R}^N, \quad N \gg 1$$

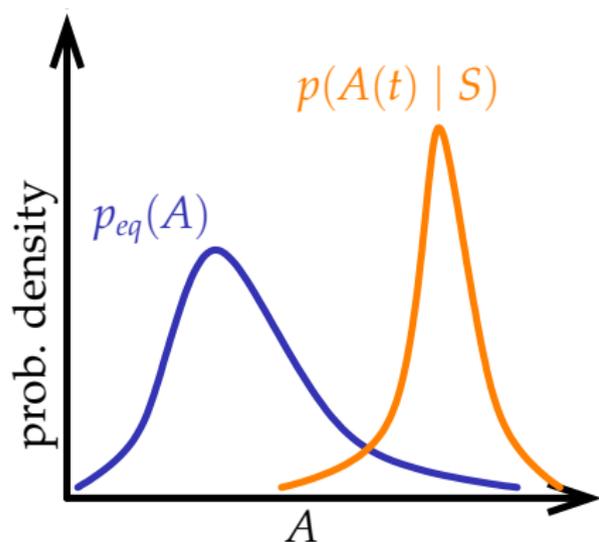
- Incomplete measurements

$$z(t) = G(x(t)), \quad z \in \mathbb{R}^n, \quad n \ll N$$

- Observable to be predicted (here assumed scalar)

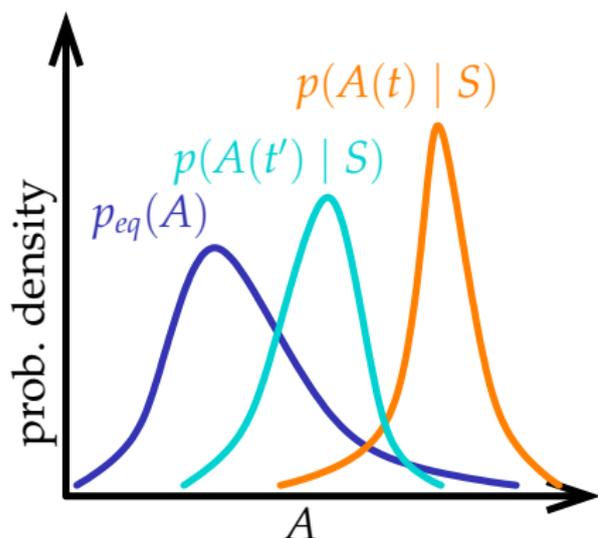
$$A(t) = A(x(t))$$

# Forecast probabilities



- The equilibrium distribution  $p_{eq}(A)$  describes our knowledge about  $A$  assuming that changes in the forcing are not important.
- $A$  is said to be *predictable* at time  $t$  if the distribution  $p(A(t) | S)$  given some initial data  $S$  differs from  $p_{eq}(A)$ .

# Forecast probabilities



- The equilibrium distribution  $p_{eq}(A)$  describes our knowledge about  $A$  assuming that changes in the forcing are not important.
- $A$  is said to be *predictable* at time  $t$  if the distribution  $p(A(t) | S)$  given some initial data  $S$  differs from  $p_{eq}(A)$ .
- For  $t'$  sufficiently larger than  $t$ ,  $p(A(t') | S)$  relaxes towards  $p_{eq}(A)$ .

Need a notion of distance between probability distributions.

What is predictability?

Revealing long-range predictability via information theory and data clustering

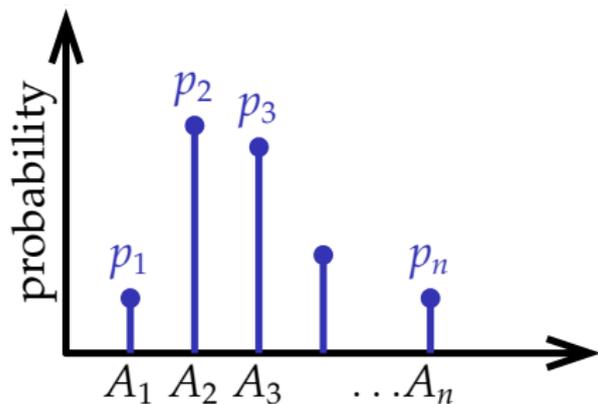
Long-range forecasts in a 1.5-layer ocean model

# Information theory



Pioneered by Claude Shannon in 1948 as a **mathematical theory of communication**.\*

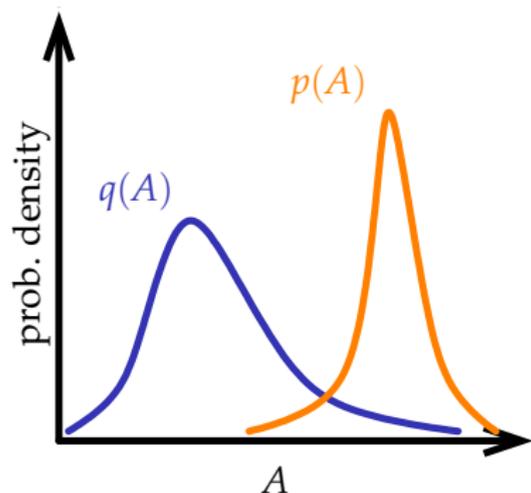
\**Bell System Technical Journal*, 27, 379, 1948.



**Entropy** measures the uncertainty about a physical system:

$$H = - \sum_i p_i \log p_i.$$

## Relative entropy



The **relative entropy** between  $p$  and  $q$  is defined as

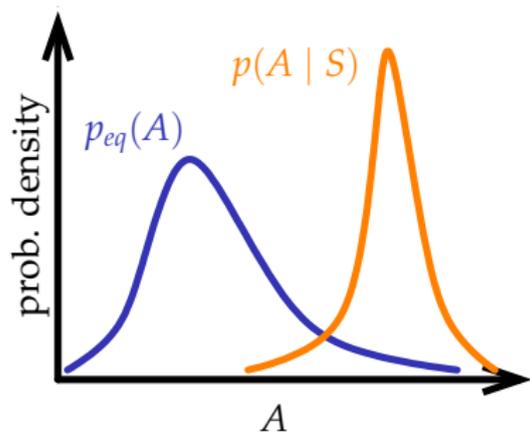
$$\mathcal{P}(p, q) = \int dA p(A) \log \frac{p(A)}{q(A)}.$$

Relative entropy describes a notion of 'distance' between probability distributions:

- $\mathcal{P}(p, q)$  is positive if  $p \neq q$ , and zero if  $p = q$ .
- $\mathcal{P}(p, q)$  is invariant under general invertible transformations of  $A$ .

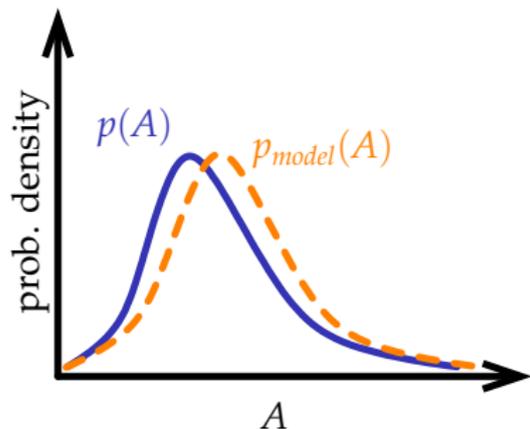
However,  $\mathcal{P}(p, q)$  is not symmetric under  $p \leftrightarrow q$ , and does not obey the triangle inequality.

# Interpretations of relative entropy



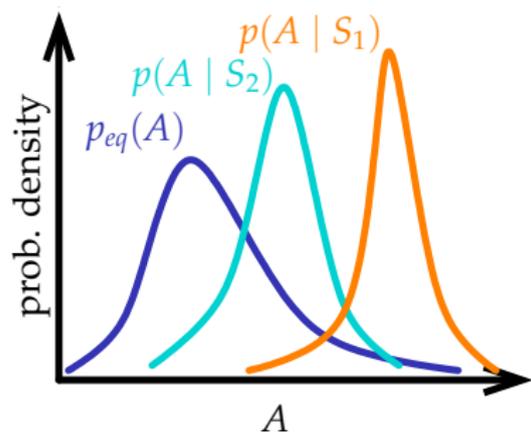
1  $\mathcal{P}(p_{posterior}, p)$  with  $p_{posterior}(A) = p(A | S)$  is the **information gain** relative to the prior distribution  $p(A)$  achieved by a measurement  $S$ .

2  $\mathcal{P}(p, p_{model})$  measures the **lack of information** or **error** in a model that describes a system by  $p_{model}(A)$  when the outcomes are actually generated by  $p(A)$ .



Note the different position of  $p$  in the 'slots' of  $\mathcal{P}(\cdot, \cdot)$ .

# Mutual information



- Different outcomes of experiments,  $S_1, S_2, \dots$ , will give rise to different amounts of information gain, as measured by  $\mathcal{P}(p(A | S_i), p(A))$ .

The expectation value of  $\mathcal{P}$  over  $S_i$  is also a relative entropy, called the **mutual information** between  $A$  and  $S$ :

$$I(A; S) = \int dS p(S) \int dA p(A | S) \log \frac{p(A | S)}{p(A)} = \mathcal{P}(p(A, S), p(A)p(S)).$$

- $I(A; S)$  is non-negative, and vanishes if and only if  $A$  and  $S$  are *statistically independent*, i.e., iff

$$p(A, S) = p(A)p(S).$$

# Fundamental questions

## Coding theory

- What is the complexity of a message  $A$ ?

*Answer:* The entropy  $H$  of the probability distribution  $p(A)$  generating the message.

- What is the ultimate rate of communication across a channel?

*Answer:* The maximum (over input distributions) of the mutual information  $I(A; B)$  between the input  $A$  and output  $B$ .

## Climate science

- What is the uncertainty in a climate variable  $A$ ?

*Answer:* The entropy  $H$  of the *equilibrium* or *climatological* distribution  $p_{eq}(A)$ .

- What is the predictability of  $A(t)$  at time  $t$  in the future, given initial data  $S$ ?

*Answer:* The mutual information  $I(A(t); S)$

# Fundamental questions

## Coding theory

- What is the expected increase in code length when a probability distribution  $q$  is used to encode a message when the message is generated by a source distribution  $p$ ?

*Answer:* The relative entropy  $\mathcal{P}(p, q)$ .

## Climate science

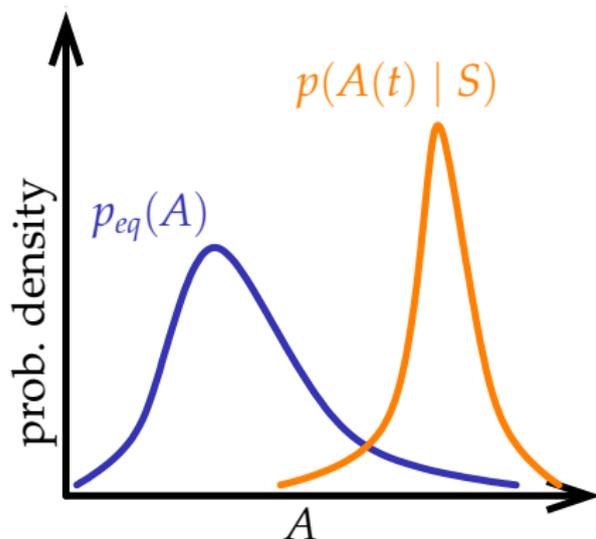
- What is the error in a model that predicts  $A(t)$  with a distribution  $p^M$  when in reality  $A(t)$  is generated by  $p$ ?

*Answer:* The relative entropy  $\mathcal{P}(p, p^M)$ .

# Quantifying predictive skill

## Relative-entropy measure of skill\*

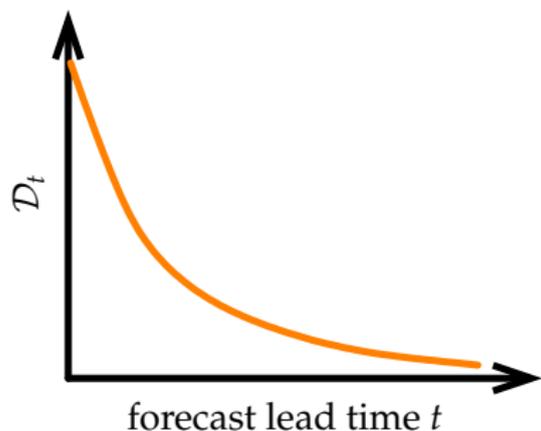
$$\mathcal{D}_t = \int dA(t) p(A(t) | S) \log \frac{p(A(t) | S)}{p_{eq}(A(t))}.$$



- $\mathcal{D}_t$  is the gain information about  $A(t)$  beyond climatology achieved by observing the initial data  $S$ .
- It is crucial that  $p(A(t) | S)$  can be computed with no *model error*. Otherwise,  $\mathcal{D}_t$  can measure false predictive skill.

\*Kleeman (2002), *J. Atmos. Sci.*, **59**, 2057; Majda et al. (2002), *Methods Appl. Anal.*, **9**, 425; DelSole (2002), *J. Atmos. Sci.*, **61**, 2425.

## Generalized second law



The relative entropy  $D_t$  is a non-increasing function of time if the following conditions are met:

- 1 As  $t \rightarrow \infty$  the conditional probabilities  $p(A(t) | S)$  converge to an equilibrium distribution  $p_{eq}(A)$  for all initial data  $S$ .
- 2  $p(A(t) | S)$  can be uniquely determined if we know  $p_0(A) = p(A(0) | S)$ .

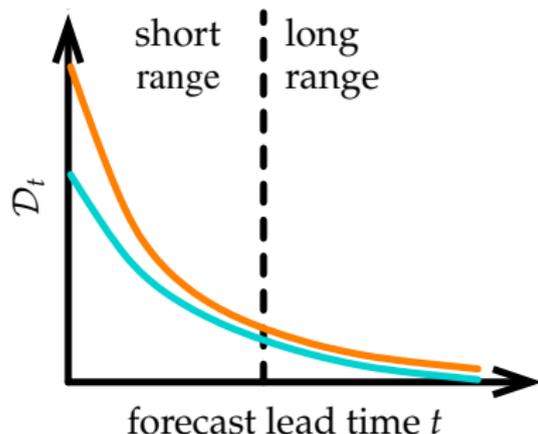
This means that the system must be *closed* and  $A(t)$  has no *memory* or *hysteresis*.

# Long-range, coarse-grained forecasts

$n$ -dim. space of initial data



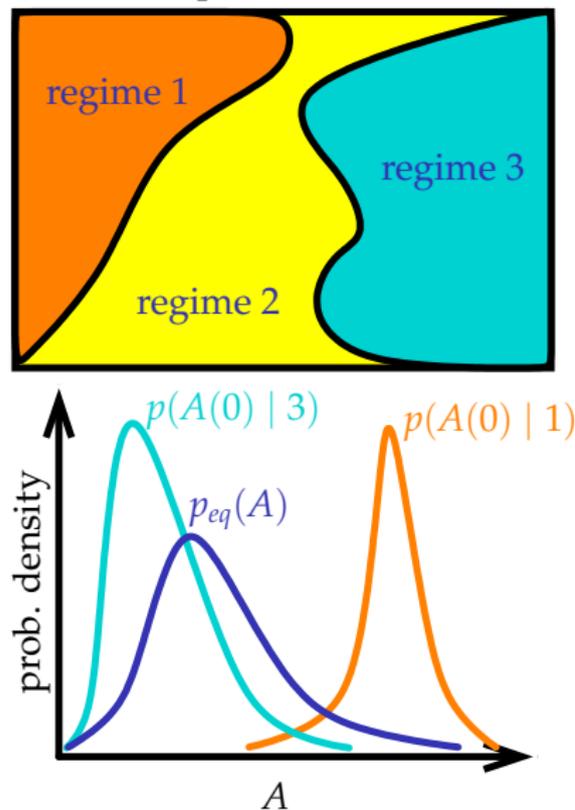
- Due to chaotic mixing, detailed initial data contribute negligible information for long-range forecasts.



- Even small uncertainties in the initial state will dominate the signal beyond a period of  $\sim$ two weeks.

# Long-range, coarse-grained forecasts

$n$ -dim. space of initial data

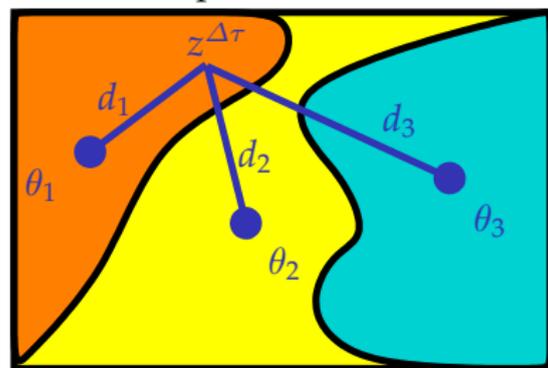


## Strategy

- 1 Instead of detailed initial data, it suffices to consider the (integer-valued) affiliation of the system to a *coarse-grained partition* of  $\mathbb{R}^n$  to make long-range forecasts.
- 2 Determine the partition empirically, by data-clustering realizations of the system in equilibrium.

# Assigning cluster affiliation

$n$ -dim. space of initial data



Each cluster is characterized by its centroid,  $\theta_k$ .

- 1 Collect observations  $z(t)$  over a time window  $\Delta\tau$  and compute the average,

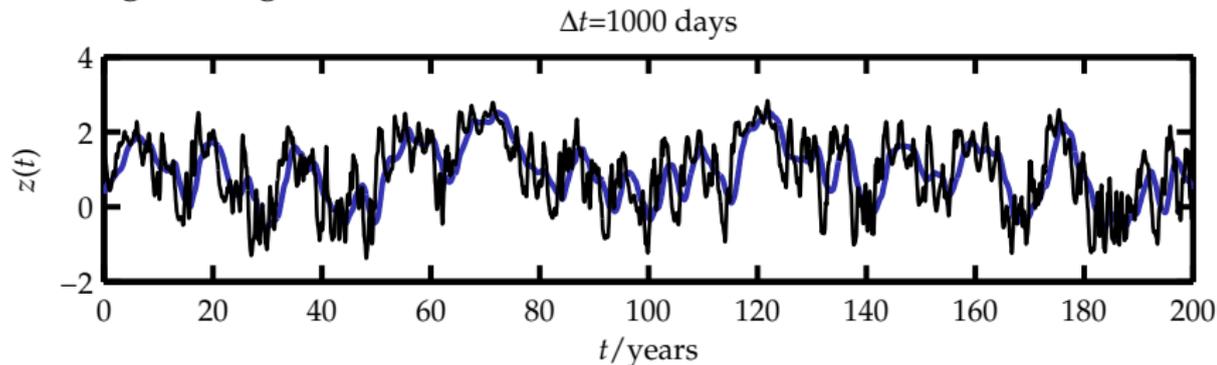
$$z^{\Delta\tau} = \frac{1}{\Delta\tau} \int_{\Delta\tau} dt z(t).$$

- 2 Set  $S$  equal to the cluster that lies closest to  $z^{\Delta\tau}$ , i.e.,

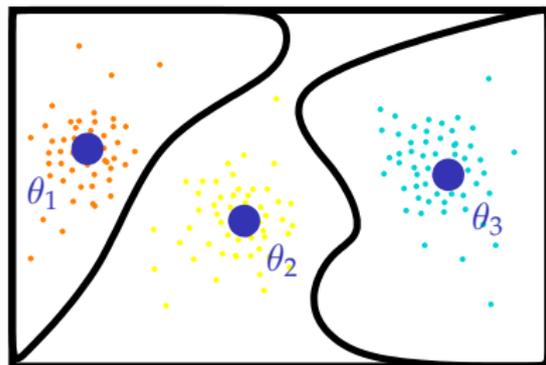
$$S = \underset{k}{\operatorname{argmin}} d_k, \quad d_k = \|z^{\Delta\tau} - \theta_k\|.$$

# Evaluating the cluster coordinates

- 1 Collect a training time series  $z(t)$ , and take the running-average over a time window  $\Delta t$ .

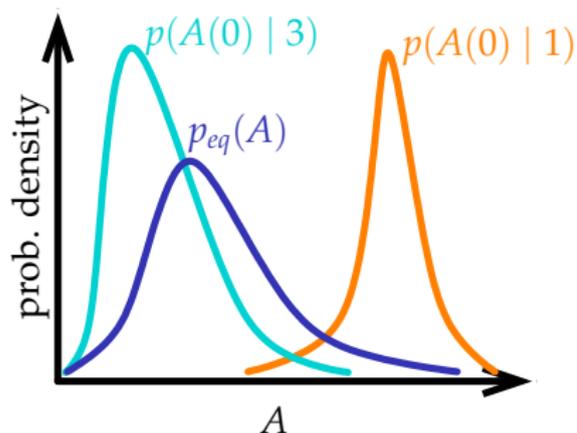


$n$ -dim space of initial data



- 2 Determine  $\theta_k$  by  $K$ -means clustering of  $z^{\Delta t}(t)$ .

# Super-ensemble forecasts



Predictive skill given that the initial data lie in the  $k$ -th coarse-grained cluster in the partition

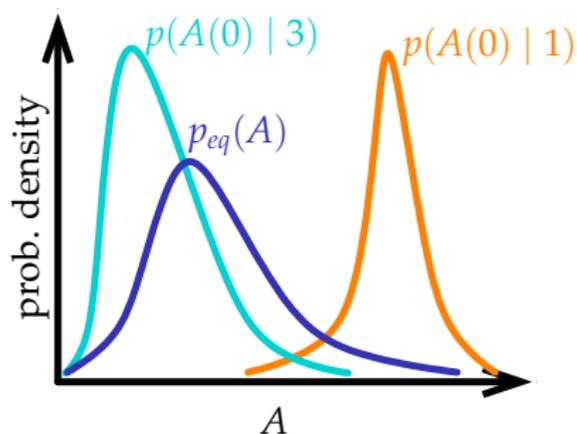
$$\mathcal{D}_t^k = \mathcal{P}(p_t^k, p_{eq}),$$

where  $p_t^k(A) = p(A(t) | S = k)$ .

The expectation value of  $\mathcal{D}_t^k$  over all cluster affiliations is a *super-ensemble* measure of predictive skill:

$$\mathcal{D}_t = \sum_{k=1}^K \pi_k \mathcal{D}_t^k, \quad \pi_k = p(S = k).$$

# Super-ensemble forecasts



Predictive skill given that the initial data lie in the  $k$ -th coarse-grained cluster in the partition

$$\mathcal{D}_t^k = \mathcal{P}(p_t^k, p_{eq}),$$

where  $p_t^k(A) = p(A(t) | S = k)$ .

## Interpretation

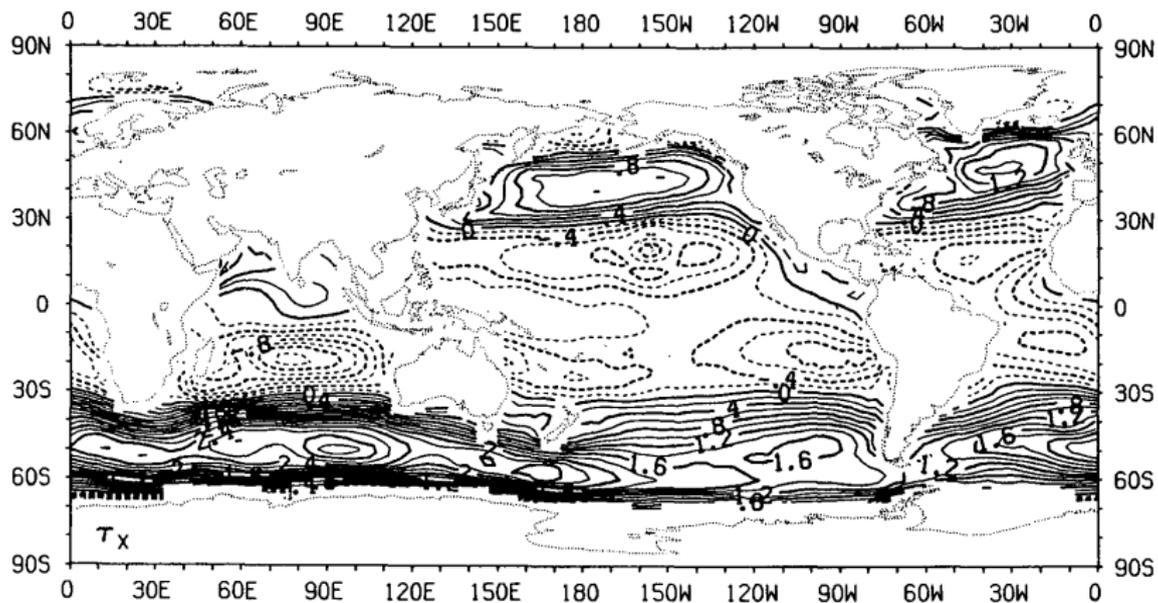
- $\mathcal{D}_t$  is equal to the mutual information  $I(A(t), S)$  between the coarse-grained initial data  $S$  and the value  $A(t)$  of the prediction observable at time  $t$ .
- $\mathcal{D}_t$  vanishes if and only if  $S$  and  $A(t)$  are statistically independent, e.g., in the  $t \rightarrow \infty$  limit.

What is predictability?

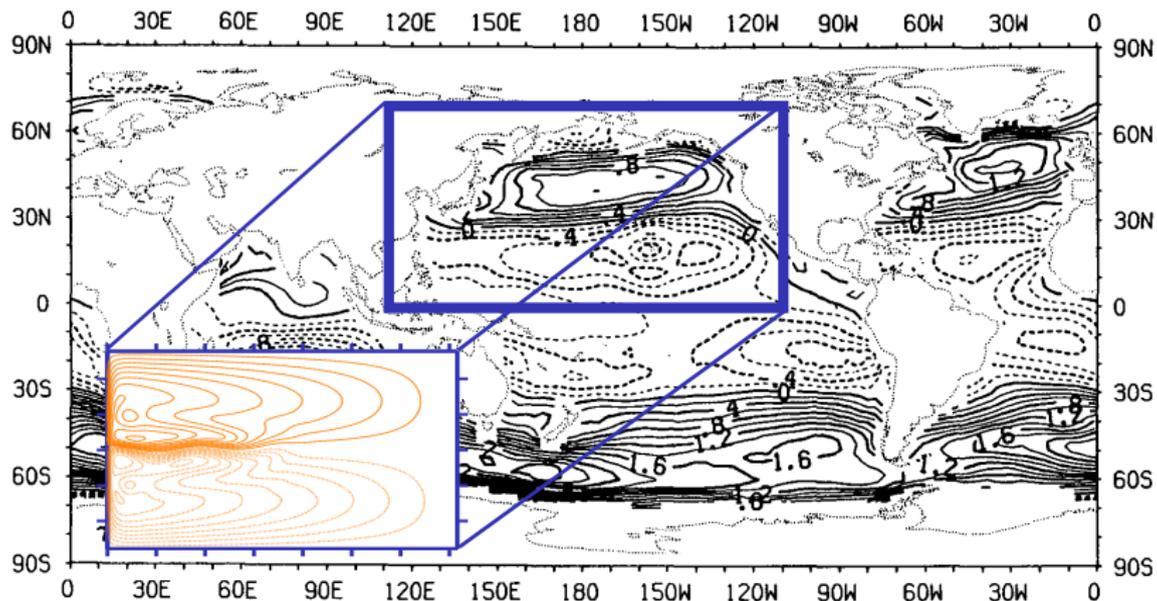
Revealing long-range predictability via information theory and data clustering

Long-range forecasts in a 1.5-layer ocean model

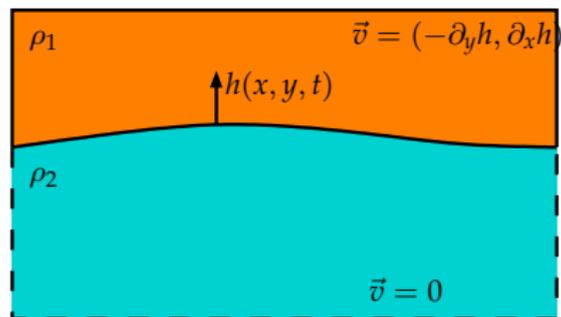
# Global ocean winds



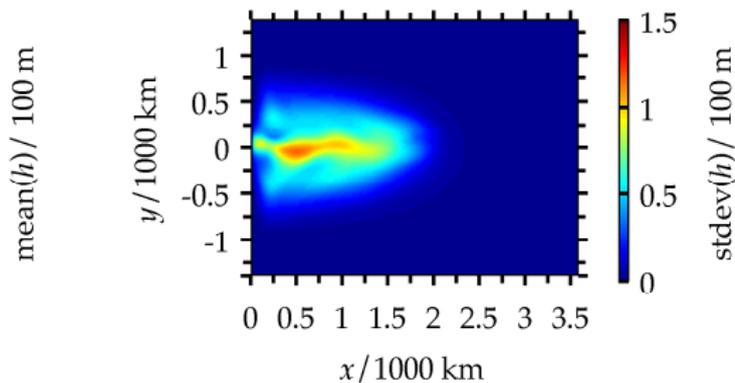
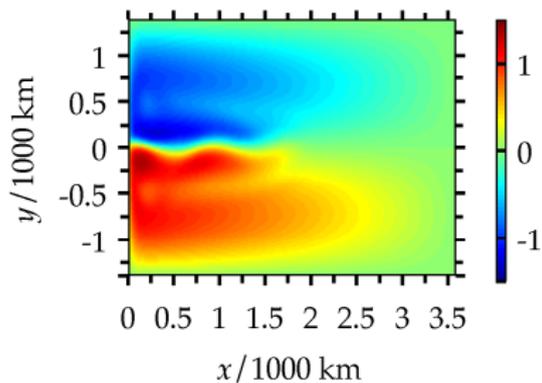
# Global ocean winds



# The 1.5-layer model<sup>†</sup>

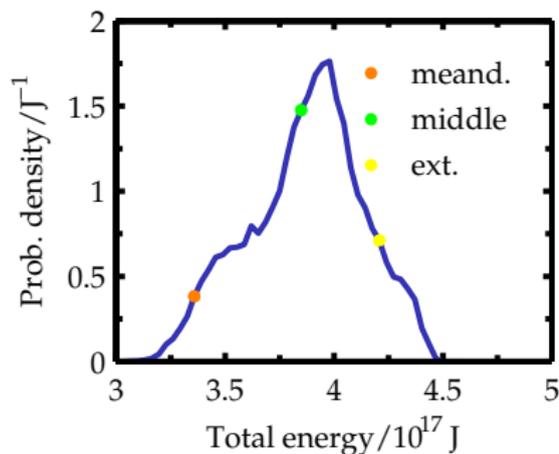
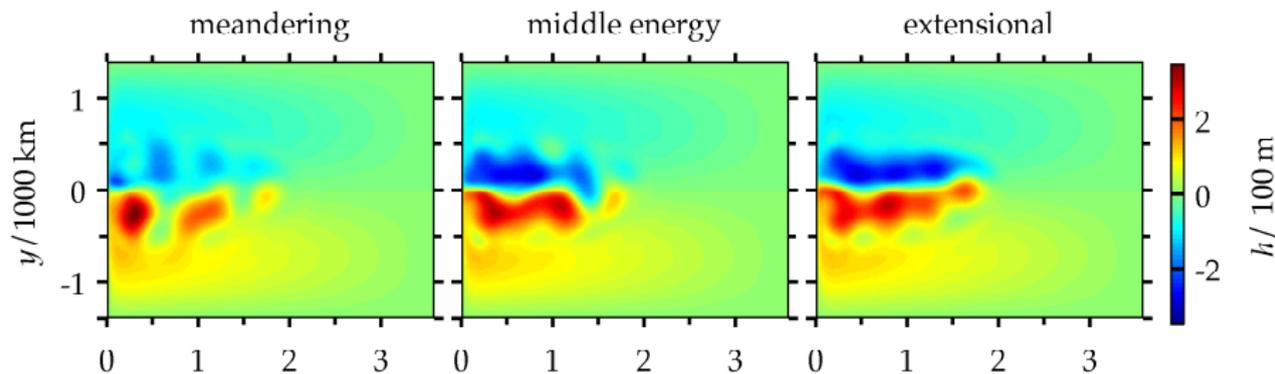


- Quasigeostrophic vorticity equation for  $h$
- Interfacial friction
- Subgrid-scale diffusion
- Asymmetric double-gyre forcing



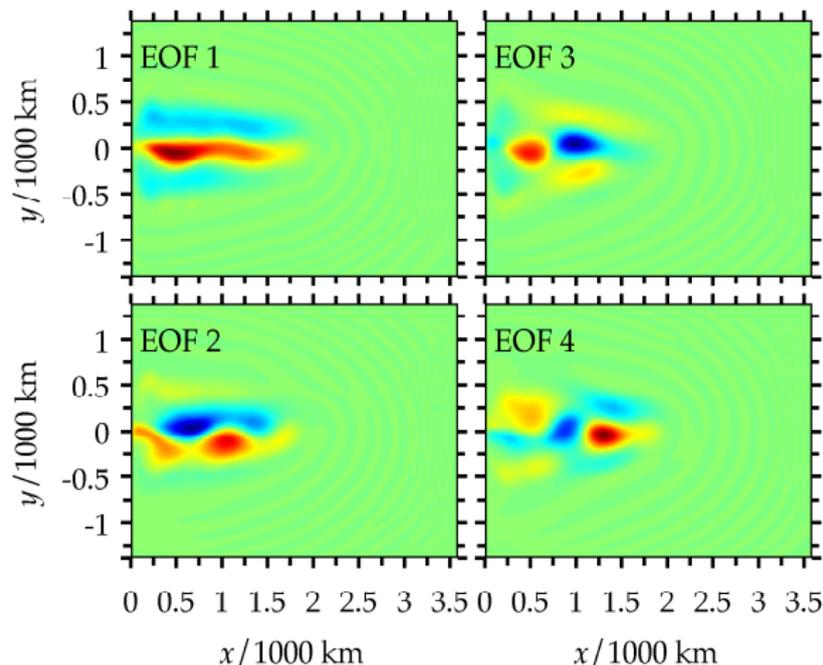
<sup>†</sup>McCalpin & Haidvogel (1996), *J. Phys. Oceanogr.*, 26, 739

# Three-state phenomenology of the 1.5-layer model



- Regimes persist for  $O(1000)$  days.
- The equilibrium distribution of energy has no local maxima.

# EOF analysis



Decompose the streamfunction in a basis of *empirical orthogonal functions (EOFs)*:

$$h(\mathbf{r}, t) = \sum_{i=1}^n \text{PC}_i(t) \text{EOF}_i(\mathbf{r}) + \text{residual.}$$

- $\text{PC}_i$  is the *principal component* corresponding to  $\text{EOF}_i$ .
- For given  $n$ , this basis minimizes the norm of the residual.

# Setup

---

Full dynamical system       $x(t) = h(\mathbf{r}, t)$

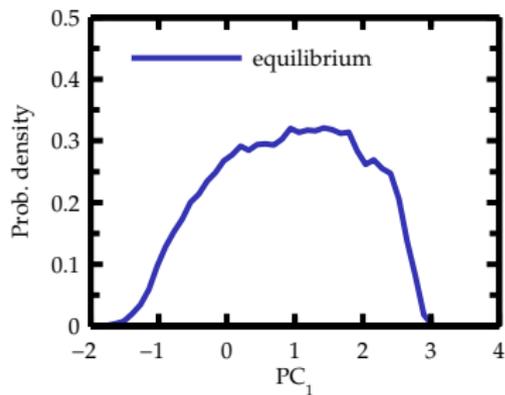
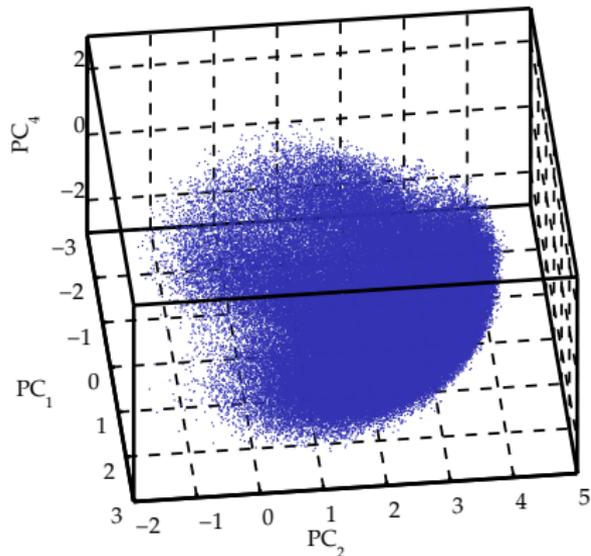
Incomplete measurements       $z(t) = (\text{PC}_1(t), \dots, \text{PC}_{20}(t))$

Prediction observables       $E$  (energy),  $\text{PC}_1, \text{PC}_2$

---

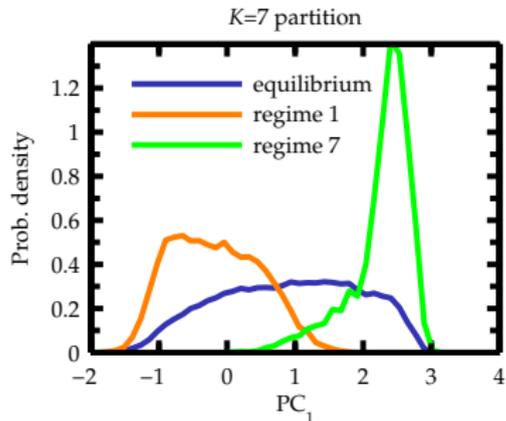
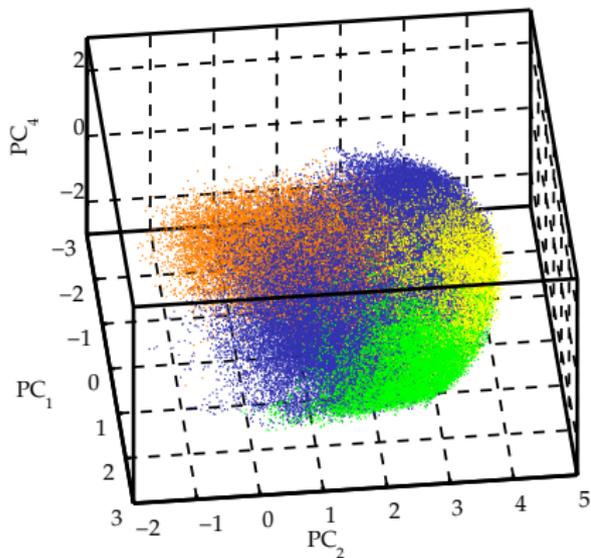
- 20-dimensional measurement vector
- Scalar prediction observables

# Coarse-grained partitions



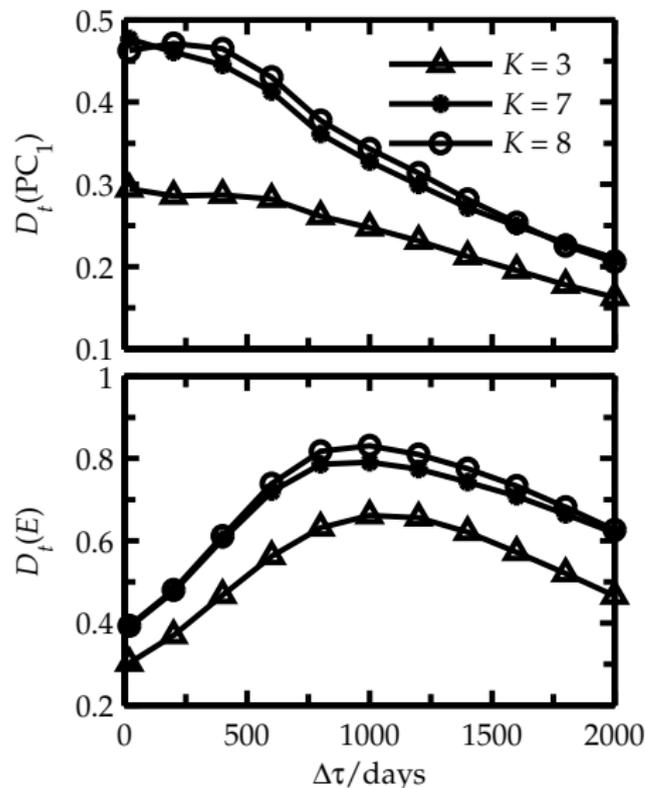
The point cloud of data in equilibrium.

# Coarse-grained partitions



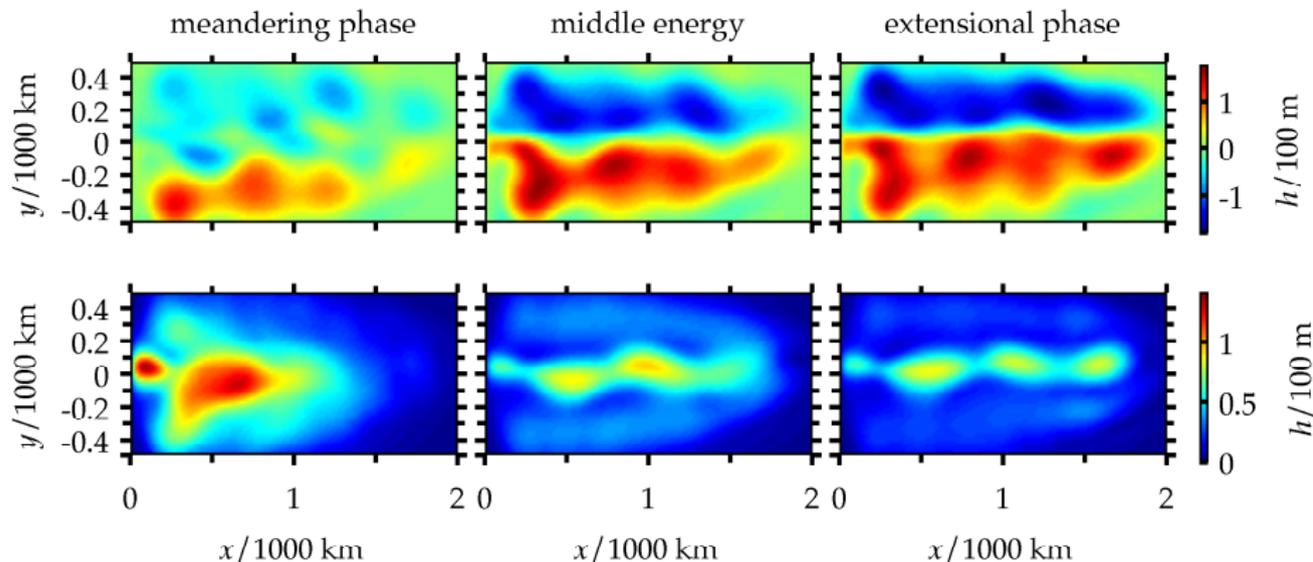
- Distinct regimes exist despite the lack of maxima in the equilibrium distributions of the PCs.
- Running-average smoothing of the training data and the initial conditions is crucial to reveal regimes.

# Information content in the partitions



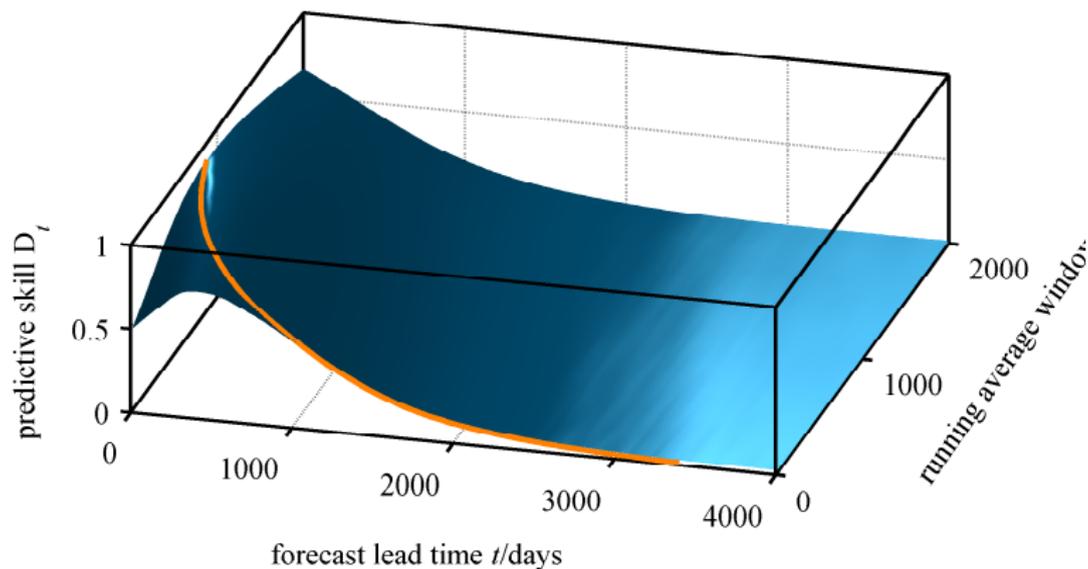
- Optimal running-average window  $\Delta\tau$  depends on the prediction observable.
  - $\Delta\tau \sim 20$  days is favored for  $PC_1$ .
  - Optimal partitions for  $E$  require more extensive coarse-graining ( $\Delta\tau \sim 1000$  days).
- The additional information content in  $K > 7$  partitions is negligible.

# Circulation regimes



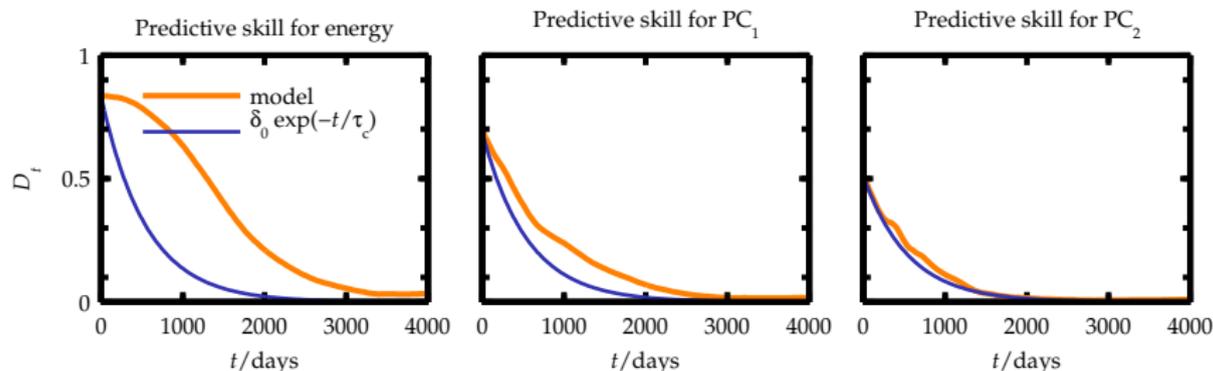
Cluster-conditional mean and standard deviation of the streamfunction anomaly.

# Long-range forecasting skill



Predictive skill for energy. The relative-entropy measure  $D_t$  reveals the optimal running-average window  $\Delta\tau$  to coarse-grain the initial data for the given prediction observable and forecast lead time.

# Long-range forecasting skill



- Predictive skill can decay more slowly than an exponential decay based on  $\tau_c$ , the longest decorrelation time of the  $z(t)$  vector used for clustering.
- Energy is predictable up to  $\sim 7$  years in advance.
- The leading PCs governing the large-scale structure of the flow are predictable up to  $\sim 5$  years in advance.

# Error in Markov models

Forecast PDFs in Markov models of regime transitions:

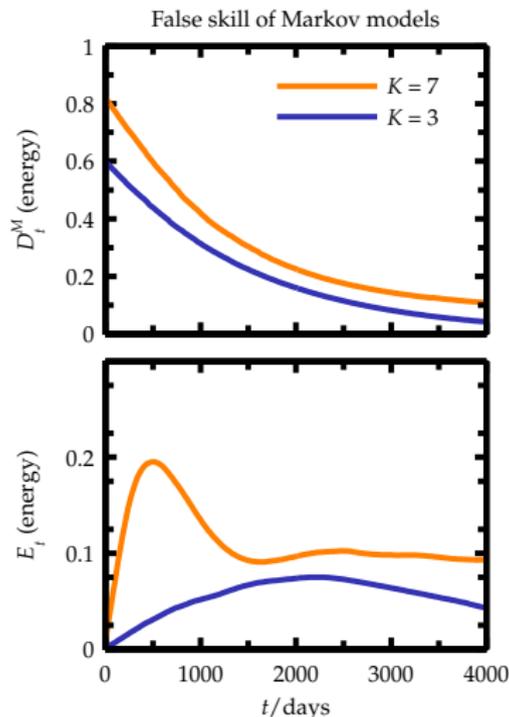
$$p_t^{Mk} = \sum_{i=1}^K [\exp(Lt)]_{ki} p_0^i.$$

Discrepancy of the Markov model from equilibrium (*can convey false skill*):

$$\mathcal{D}_t^k = \int p_t^{Mk} \log \frac{p_t^{Mk}}{p_{eq}^M}.$$

Error in the Markov model relative to the true model:

$$\mathcal{E}_t^k = \int p_t^k \log \frac{p_t^k}{p_t^{Mk}}.$$



- Equilibrium consistency,  $\mathcal{E}_\infty^k \ll 1$ , is essential for long-range forecasts.
- Persistence of imperfect models is not synonymous with fidelity.

## Conclusions & outlook

- Information theory provides an objective measure of predictive skill as the **relative entropy** between the distribution  $p(A(t) | S)$  conditioned on initial data  $S$  and the equilibrium distribution  $p_{eq}(A)$ .
- A natural measure for **model error** is the relative entropy between  $p(A(t) | S)$  and the forecast distribution  $p^M(A(t) | S)$  in an imperfect model.
- For long-range forecasts it suffices to take  $S$  to be the affiliation of the system at time  $t = 0$  to a **coarse-grained partition** of the set of possible initial data.
  - No need to specify detailed initial data (a major challenge in ensemble forecasts).

## Conclusions & outlook

- Applied in a simple double-gyre ocean model, the technique reveals
  - Circulation regimes are consistent with empirical phenomenology of rolled-up and extensional configurations of the jet.
  - Predictability in certain large-scale observables approaching 10 years.
- Error in Markov models of the low-frequency dynamics can be assessed *a posteriori*.
- Currently working on extensions of these methods to comprehensive climate models featuring baroclinic effects and time-dependent statistics.

## References

- Kleeman (2002), Measuring dynamical prediction utility using relative entropy, *J. Atmos. Sci.*, **59**, 2057.
- Majda, Abramov & Grote (2005), *Information Theory and Stochastics for Multiscale Nonlinear Systems*, CRM Monograph Series, Vol. 25, AMS.
- DelSole & Tippett (2007), Predictability: Recent insights from information theory, *Rev. Geophys.*, **45**, RG4002.
- Meehl et al. (2009), Decadal prediction. Can it be skillful?, *Bull. Amer. Meteor. Soc.*, **90**, 1467.
- Majda & Gershgorin (2010), Quantifying uncertainty in climate change science through empirical information theory, *Proc. Natl. Acad. Sci.*, **107**, 14958.

[this work]

- Giannakis & Majda (2011), Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model, submitted to *J. Climate*.
- Giannakis & Majda (2011), Quantifying the predictive skill in long-range forecasting. Part II: Model error in coarse-grained Markov models with application to ocean-circulation regimes, submitted to *J. Climate*.