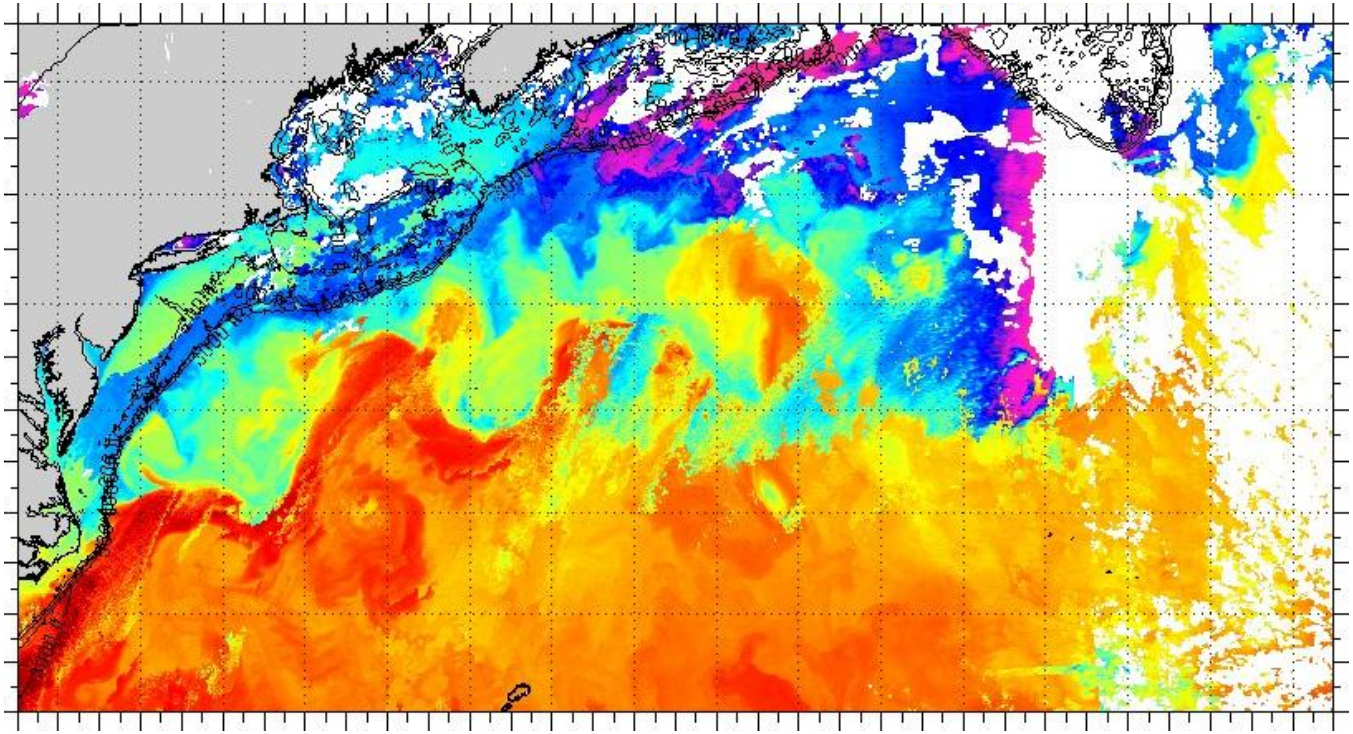
Quantifying predictability and model error in long-range climate forecasting through information theory



Dimitris Giannakis
Courant Institute of Mathematical Sciences

ICiS Workshop on VVUQ, August 9 2011

Outline

1. What is predictability?
   - Initial-value vs. forced-response problem.
2. Information theory and climate forecasts.
   - Predictability as information gain beyond equilibrium climate.
   - Model error as lack of information relative to nature.
3. Long-range forecasts in a double-gyre ocean model.
   - Revealing long-range predictability by data clustering.
   - Model error in Markov models of regime transitions.

Key points

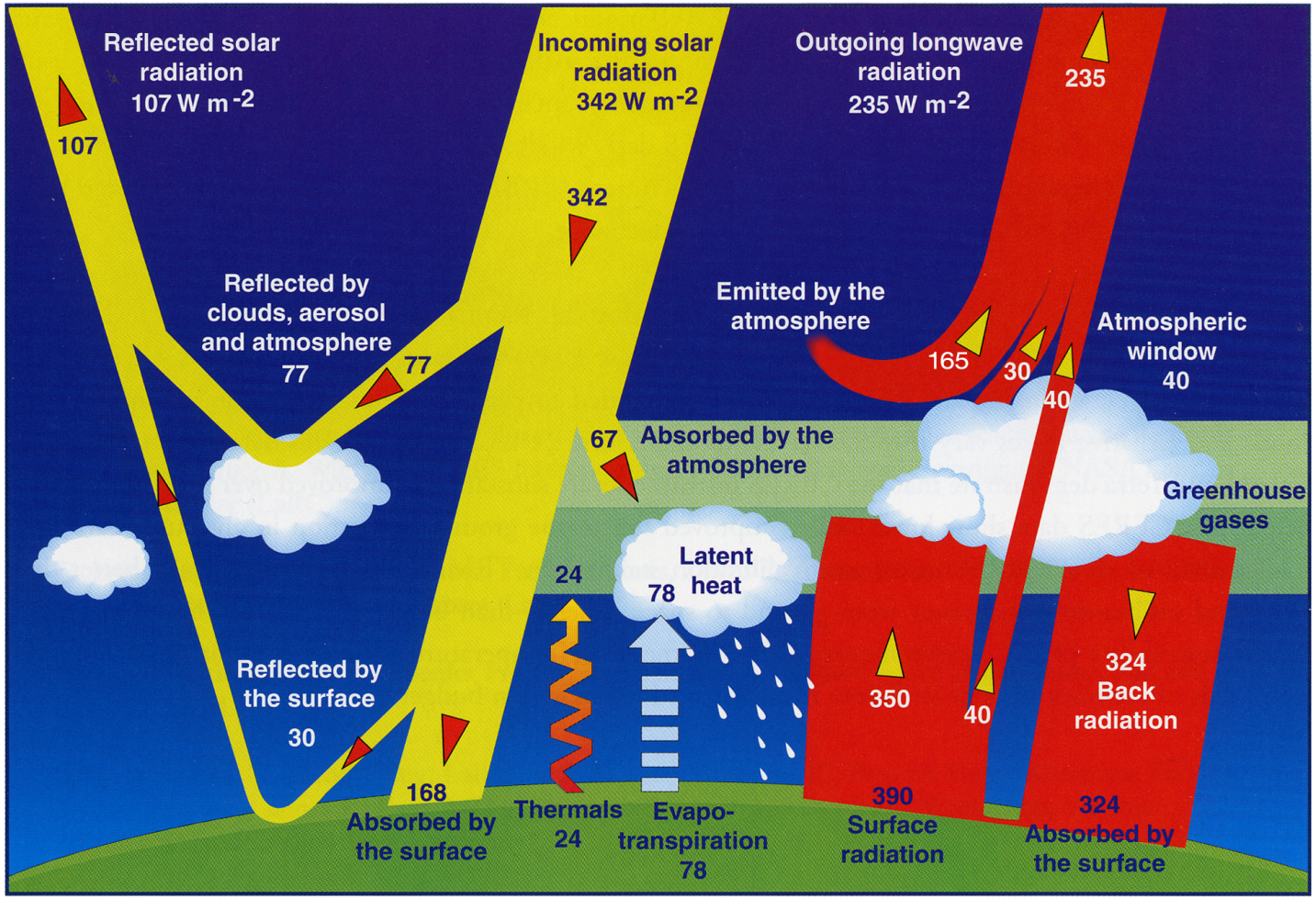
- Dynamical-system predictions as communication across a channel.
  - Predictability $\Longleftrightarrow$ channel capacity.
  - Model error $\Longleftrightarrow$ suboptimal coding.
- Low-frequency variability does not necessarily imply predictability.
  - Identifying predictability via relative entropy.
- Long-range predictability can be revealed by an appropriate coarse-grained partition of a set of (incomplete) initial data.
  - Clustering helps beat the "curse of dimension".
- Forecasts with imperfect models must be assessed via
  1. Fidelity in reproducing the equilibrium climate.
  2. Dynamical error during relaxation to equilibrium.

Acknowledgments

Joint work with Andrew Majda

Discussions	Boris Gershgorin (Courant) Illia Horenko (Univ. Lugano)
Dataset	Rafail Abramov (Univ. Illinois Chicago)

The climate's energy budget



Kiehl & Trenberth (1997), *Bull. Amer. Met. Soc.*, **78**, 197.

Initial-value vs. forced-boundary forecasts

lead time →

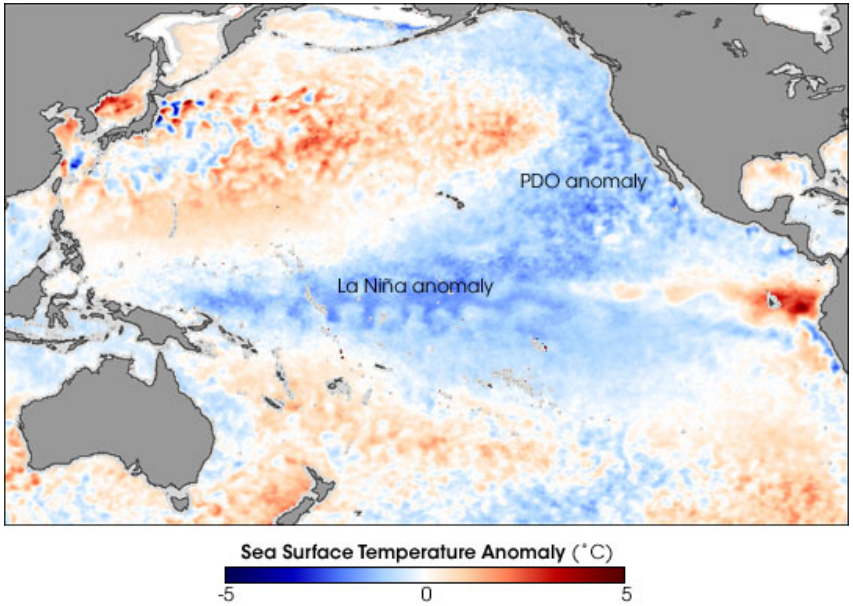
daily weather forecasts	seasonal outlooks	decadal predictions	century climate projections
initial-value problem			
			forced boundary-value problem

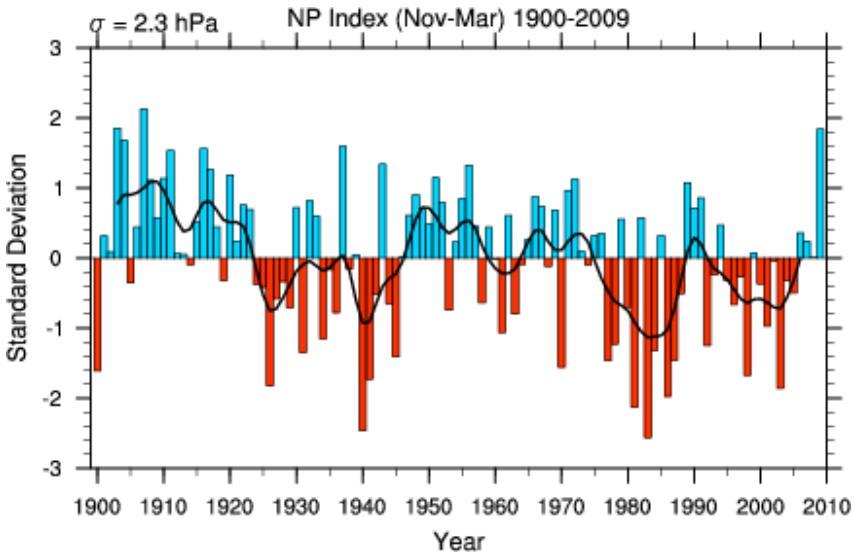
↑

regime behavior
North Atlantic oscillation,
Pacific Decadal oscillation, …

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

Pacific decadal oscillation (PDO)

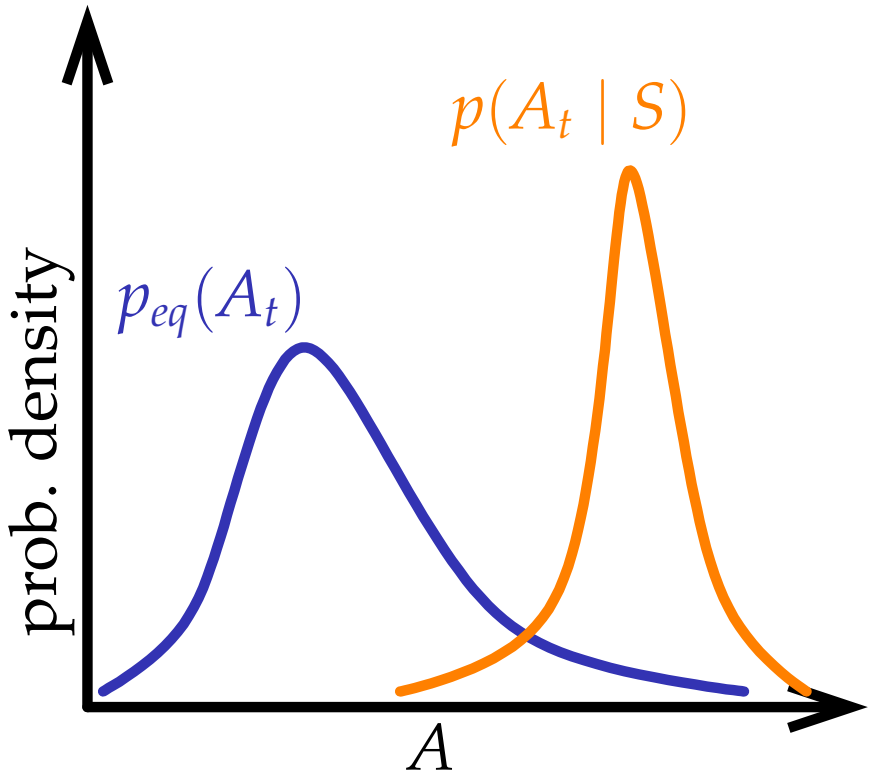




- A low-frequency oscillation of the sea surface temperature in the North Pacific, classified as being in either warm or cool phases.
- Cool phase (shown left) is associated with below-average precipitation in the Southwestern US.
- Only 2–3 cycles in the last century.
- Dynamical origins and predictability are an area of active research.

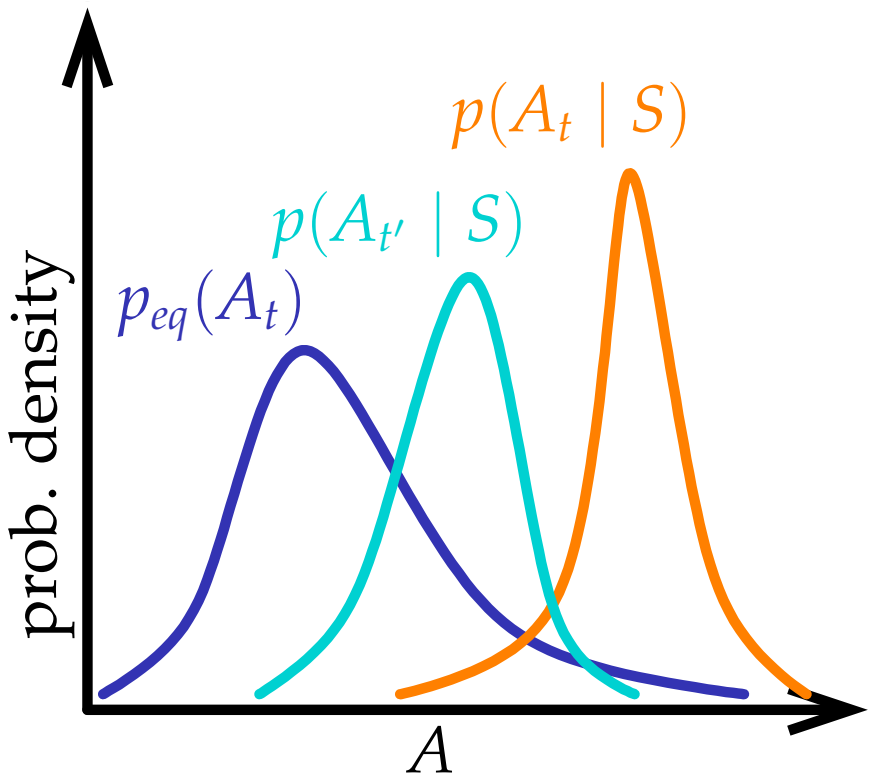
Image sources: `earthobservatory.nasa.gov/IOTD`, `www.cgd.ucar.edu/cas/jhurrell/indices.info.html`

Forecast probabilities



- The equilibrium distribution $p_{eq}(A_t)$ describes our prior knowledge about an observable A_t, assuming that changes in external forcings are not important.
- A_t is said to be *predictable* at time t if the distribution $p(A_t \mid S)$ given some initial data S differs from $p_{eq}(A_t)$.

Forecast probabilities



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

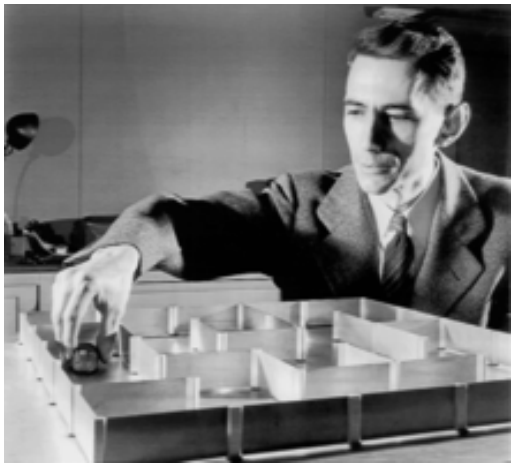
Need a notion of distance between probability distributions.

What is predictability?

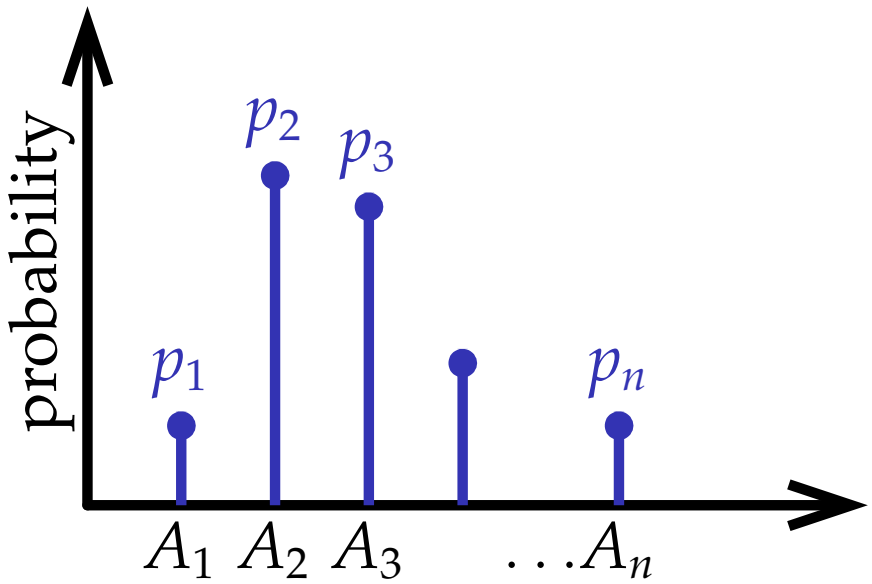
Information theory and climate predictions

Long-range forecasts in a double-gyre ocean model

Information theory



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

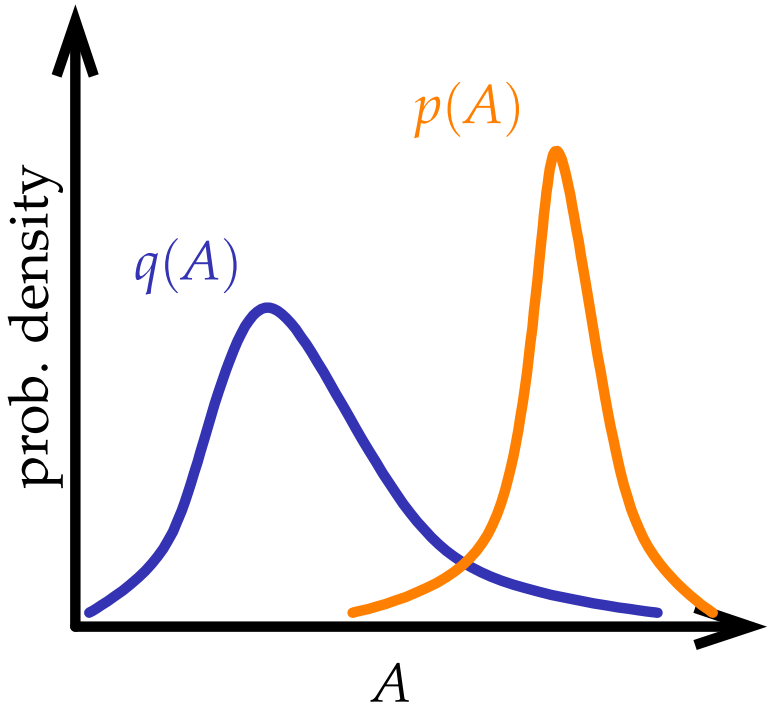


Entropy measures the uncertainty about a physical system:

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

*Shannon (1948), *Bell System Technical Journal*, **27**, 379.

Relative entropy



Relative entropy:

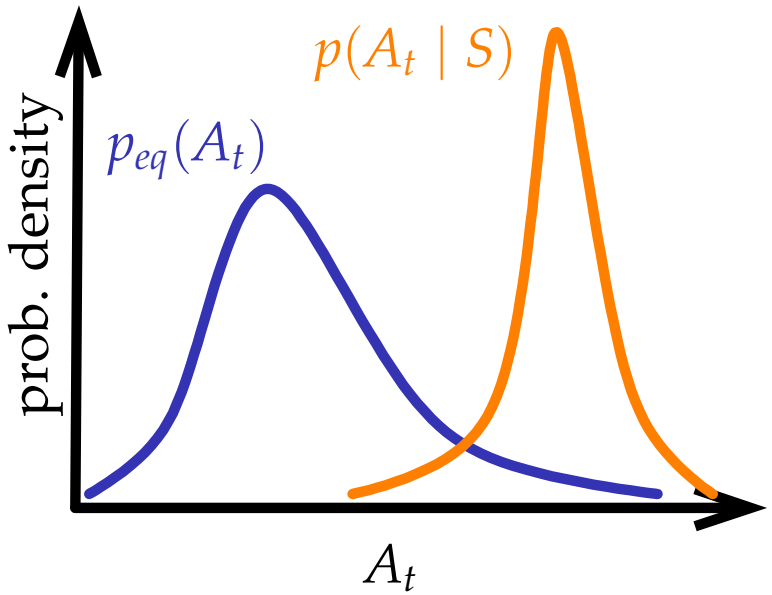
$$\mathcal{P}(p,q) = \int dA\, p(A) \log \frac{p(A)}{q(A)}.$$

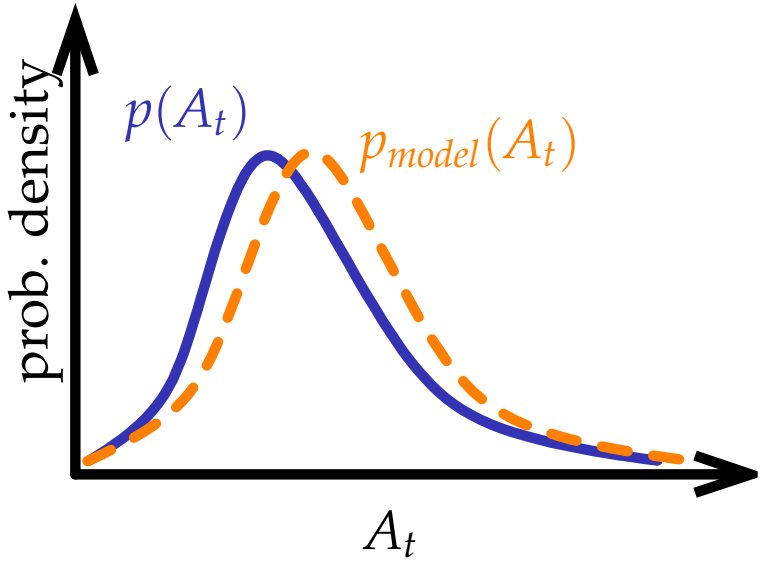
Relative entropy describes a notion of "distance" between probability distributions:

- $\mathcal{P}(p,q)$ 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

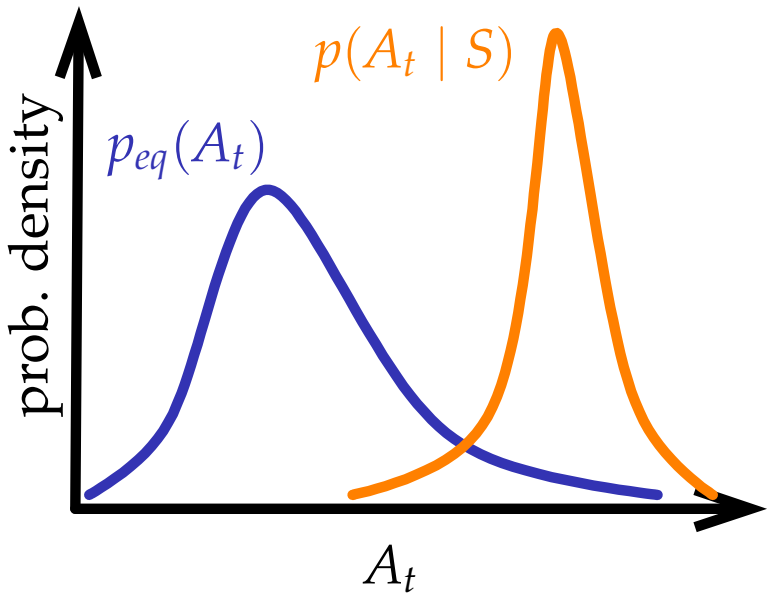


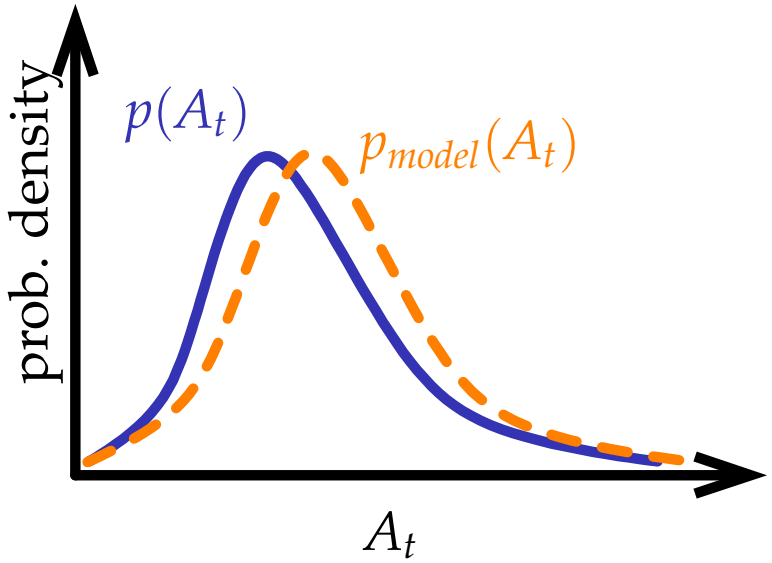


1. $\mathcal{P}(p_{posterior}, p)$ with $p_{posterior} = p(A_t \mid S)$ is the *information gain* relative to the prior distribution $p(A)$ achieved by a measurement S.
2. $\mathcal{P}(p, p_{model})$ measures the *lack of information* or *ignorance* in a distribution that describes a system by p_{model} when the outcomes are actually generated by p.

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

Quantifying predictability and model error*





- *Predictability:* Gain of information about A_t beyond equilibrium achieved by observing the initial data S:

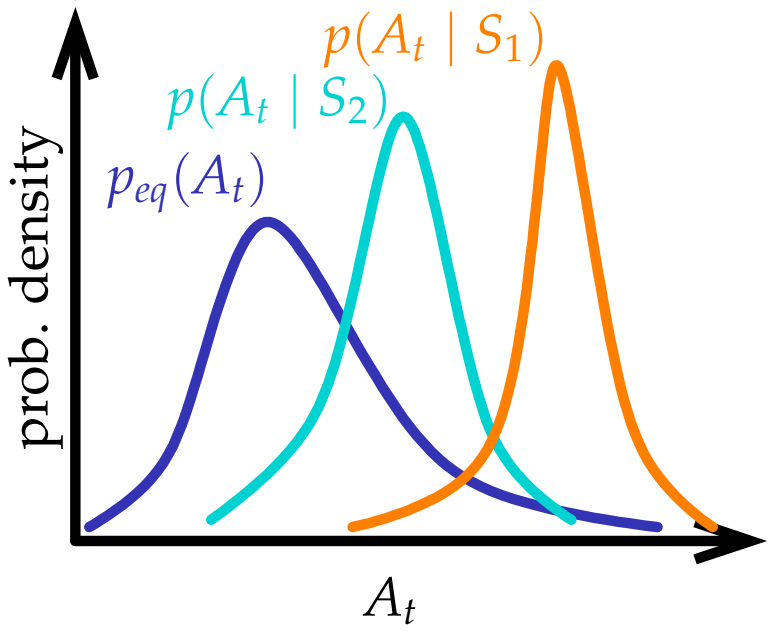
$$\mathcal{D}_t = \mathcal{P}(p(A_t \mid S), p_{eq}(A_t)).$$

- *Model error:* Lack of information in the model probability distribution p^M relative to the truth:

$$\mathcal{E}_t = \mathcal{P}(p(A_t \mid S), p^M(A_t \mid S)).$$

*Kleeman (2002), *J. Atmos. Sci.*, **59**, 2057; Majda et al. (2002), *Methods Appl. Anal.*, **9**, 425; Roulston & Smith (2002), *Mon. Weather Rev.*, **130**, 1653.

Mutual information



- Different initial data, $S_1, S_2, \ldots$, will give rise to different amounts of information gain, as measured by $\mathcal{P}(p(A_t \mid S_i), p(A_t))$.

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

$$I(A_t; S) = \int dS\, p(S) \int dA_t\, p(A_t \mid S) \log \frac{p(A_t \mid S)}{p(A_t)} = \mathcal{P}(p(A_t, S), p(A_t)p(S)).$$

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

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

Fundamental questions

Coding theory

- 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.

- 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

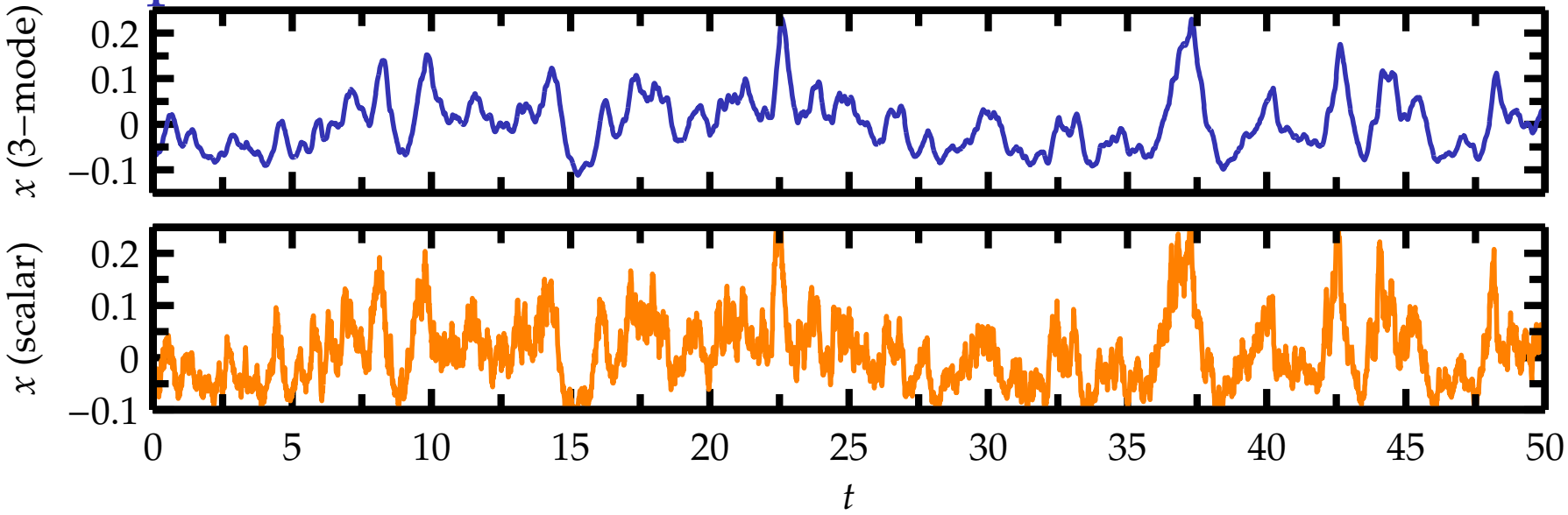
- What is the predictability of A_t at time t in the future, given initial data S?

  Answer: The mutual information $I(A_t;S)$

- 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 $\mathcal{P}(p,p^M)$.

Example: Three-mode stochastic model



- Three-mode model with linear and dyad interactions between mode x (resolved) and y_i (unresolved):

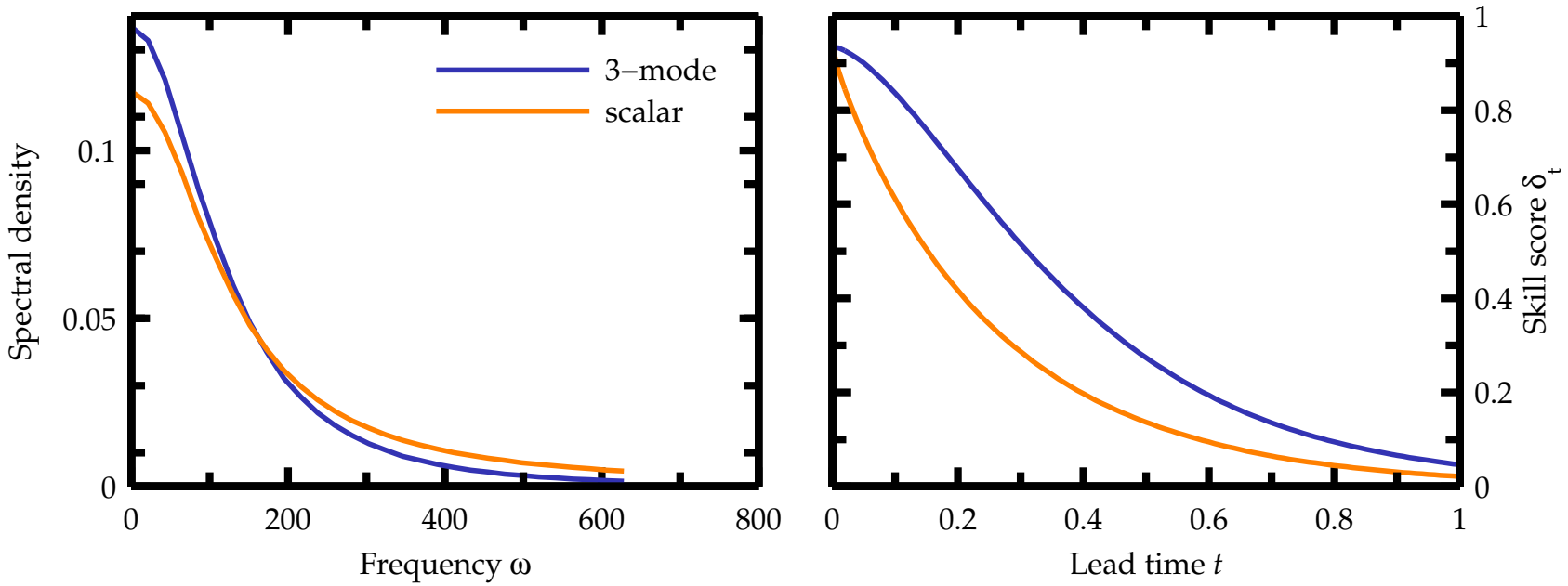
$$
\begin{aligned}
dx &= (Ixy_1 + L_1 y_1 + L_2 y_2 + F + Dx)\, dt \\
dy_1 &= \left(-Ix^2 - L_1 x - \gamma_1 \epsilon^{-1} y_1\right)\, dt + \sigma_1 \epsilon^{-1/2}\, dW_1, \\
dy_2 &= \left(-L_2 x - \gamma_2 \epsilon^{-1} y_2\right)\, dt + \sigma_2 \epsilon^{-1/2}\, dW_2.
\end{aligned}
$$

- Reduced scalar model, describing the limit $\epsilon \to 0$:*

$$
dx = (\tilde{F} + ax + bx^2 - cx^3)\, dt + (\alpha - \beta x)\, dW_1 + \sigma\, dW_2.
$$

*Majda, Timofeyev, Vanden Eijnden (2003), *J. Atmos.Sci.*, **60**, 1705.

Example: Three-mode stochastic model



- For small-enough ϵ (e.g., $\epsilon = 0.1$ shown here), the three-mode and scalar models have comparable low-frequency variability.
- Measure predictability via

$$\mathcal{D}_t = I(x(t); x_0), \quad \delta_t = 1 - \exp(-2\mathcal{D}_t).$$

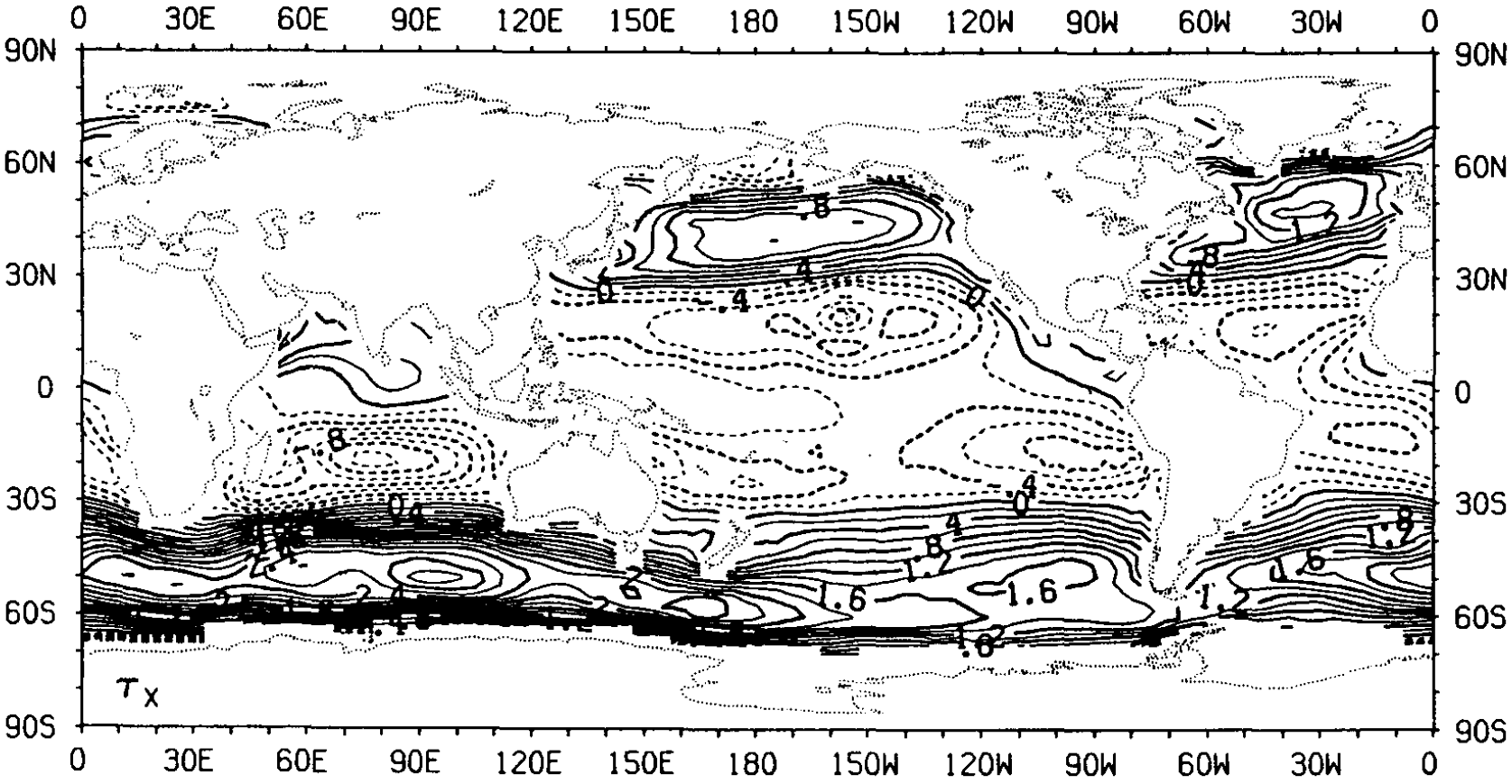
- Relative entropy reveals the loss of predictability due to the stochastic approximation of the interaction between x and (y_1, y_2).

What is predictability?

Information theory and climate predictions

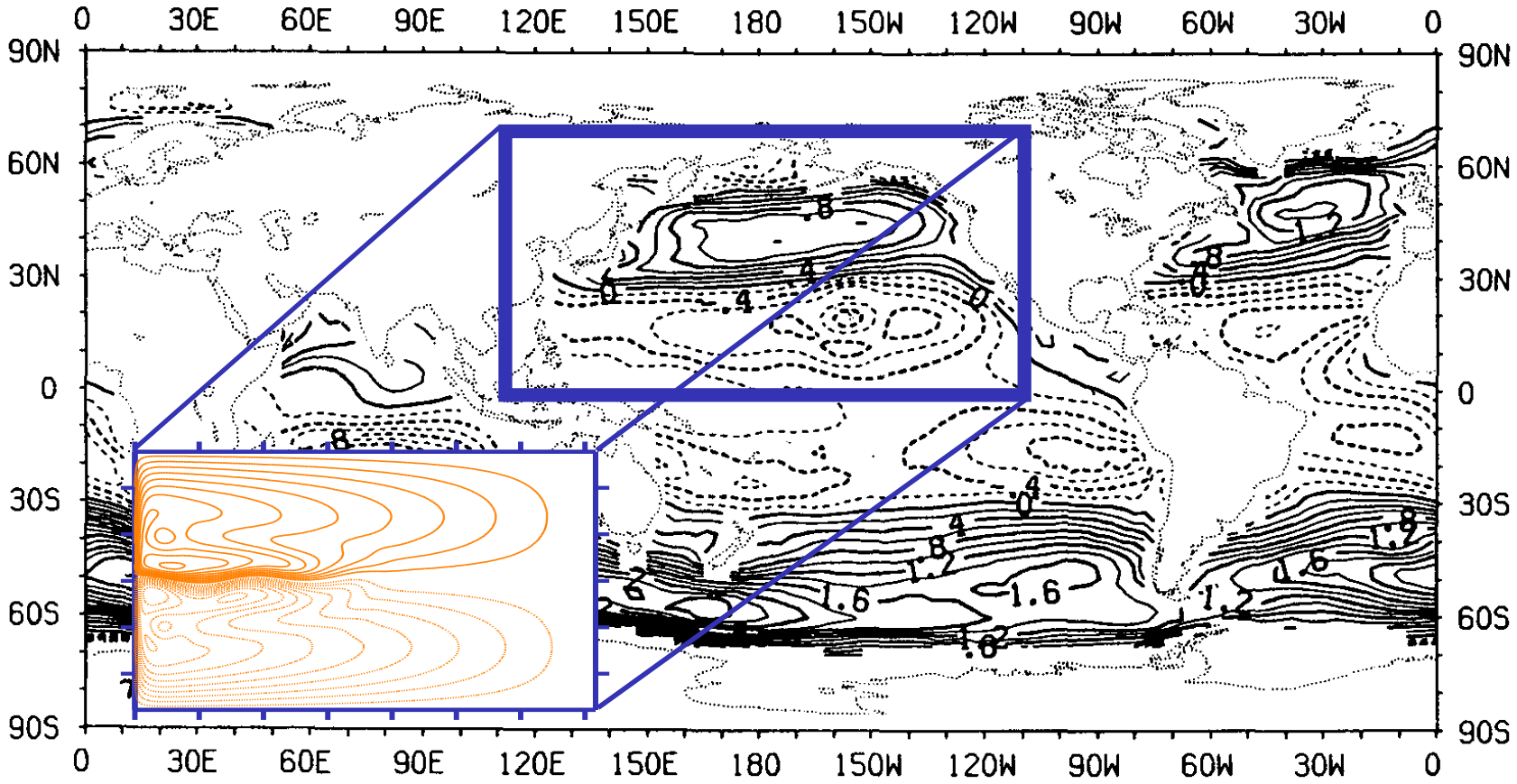
Long-range forecasts in a double-gyre ocean model

Global ocean winds



Source: Trenberth *et al.* (1990), *J. Phys. Oceanogr.*, **20**, 1742.

Global ocean winds

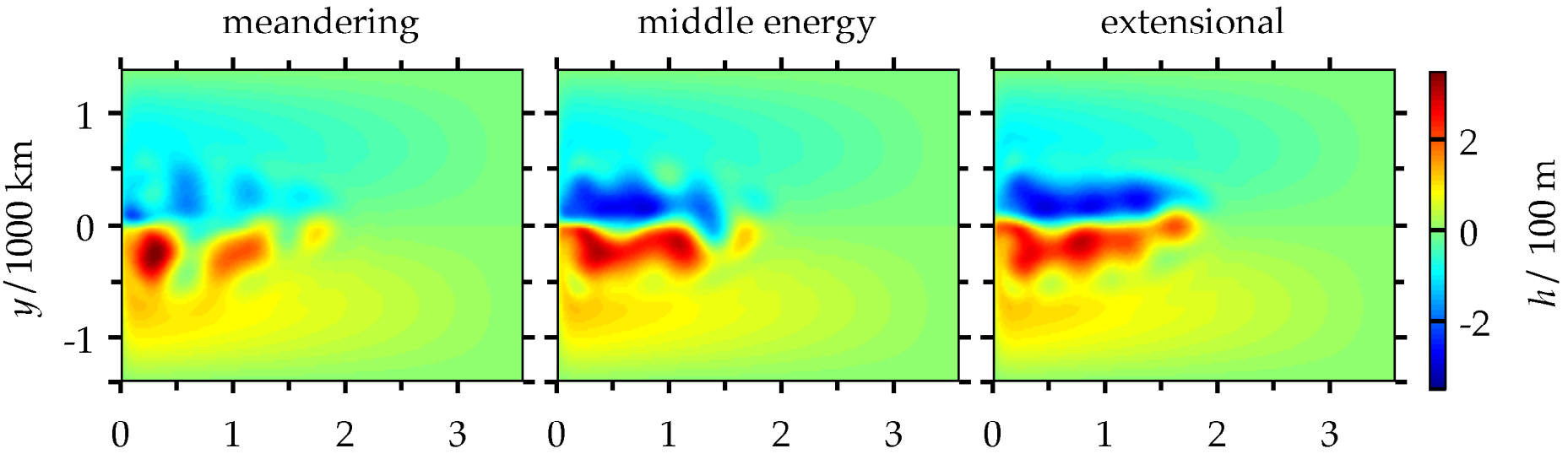


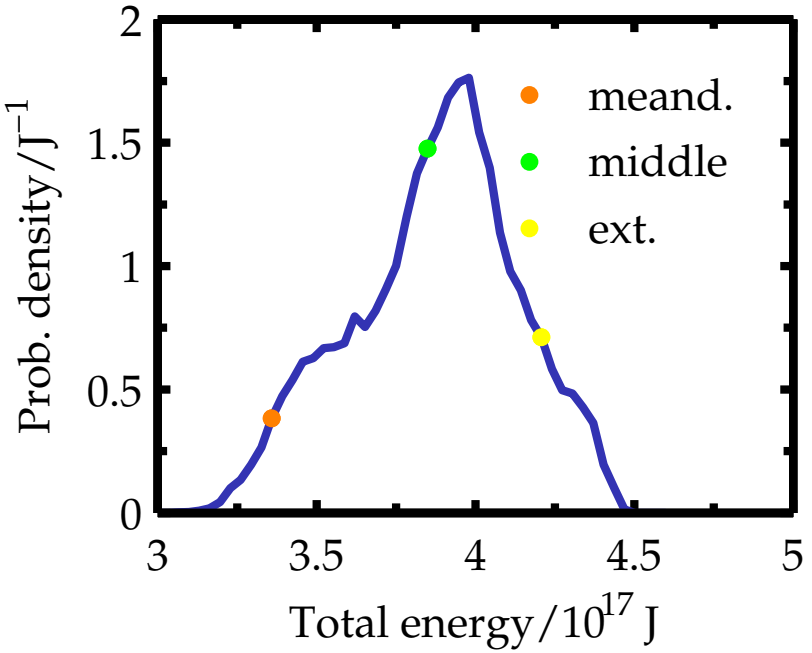
Source: Trenberth *et al.* (1990), *J. Phys. Oceanogr.*, **20**, 1742.

Typical ocean basin simulated using a "1.5-layer" model with constant wind forcing.*

*McCalpin & Haidvogel (1996), *J. Phys. Oceanogr.*, 26, 739.

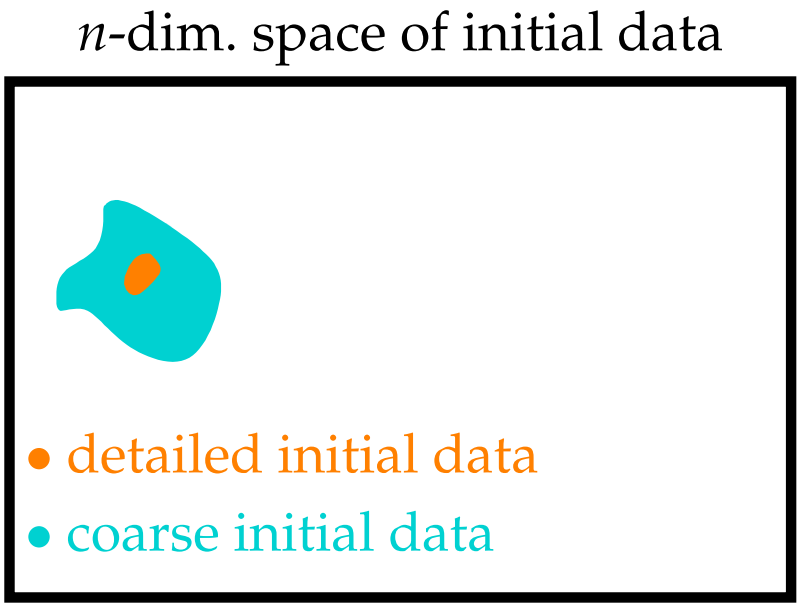
Three-state phenomenology of the 1.5-layer model

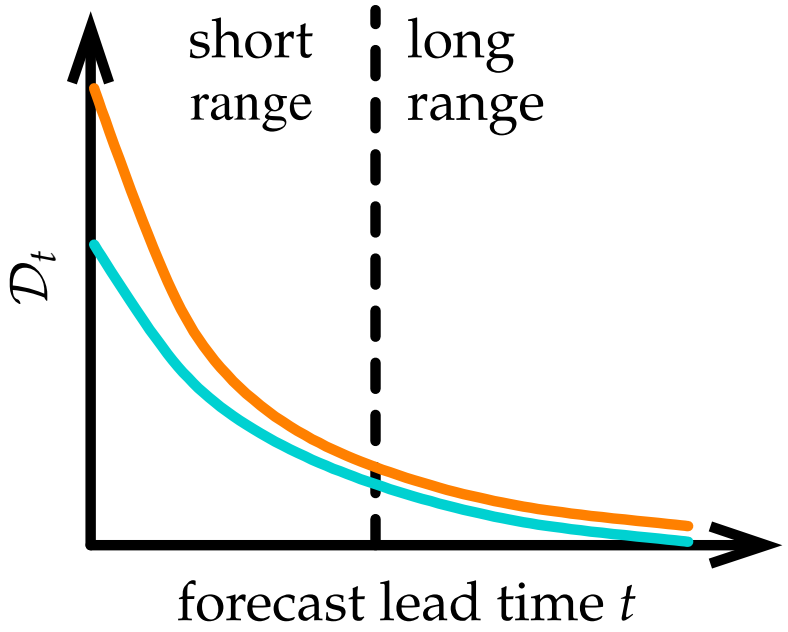




- Regimes persist for O(1000) days.
- The equilibrium distribution of energy has no local maxima.

Long-range, coarse-grained forecasts

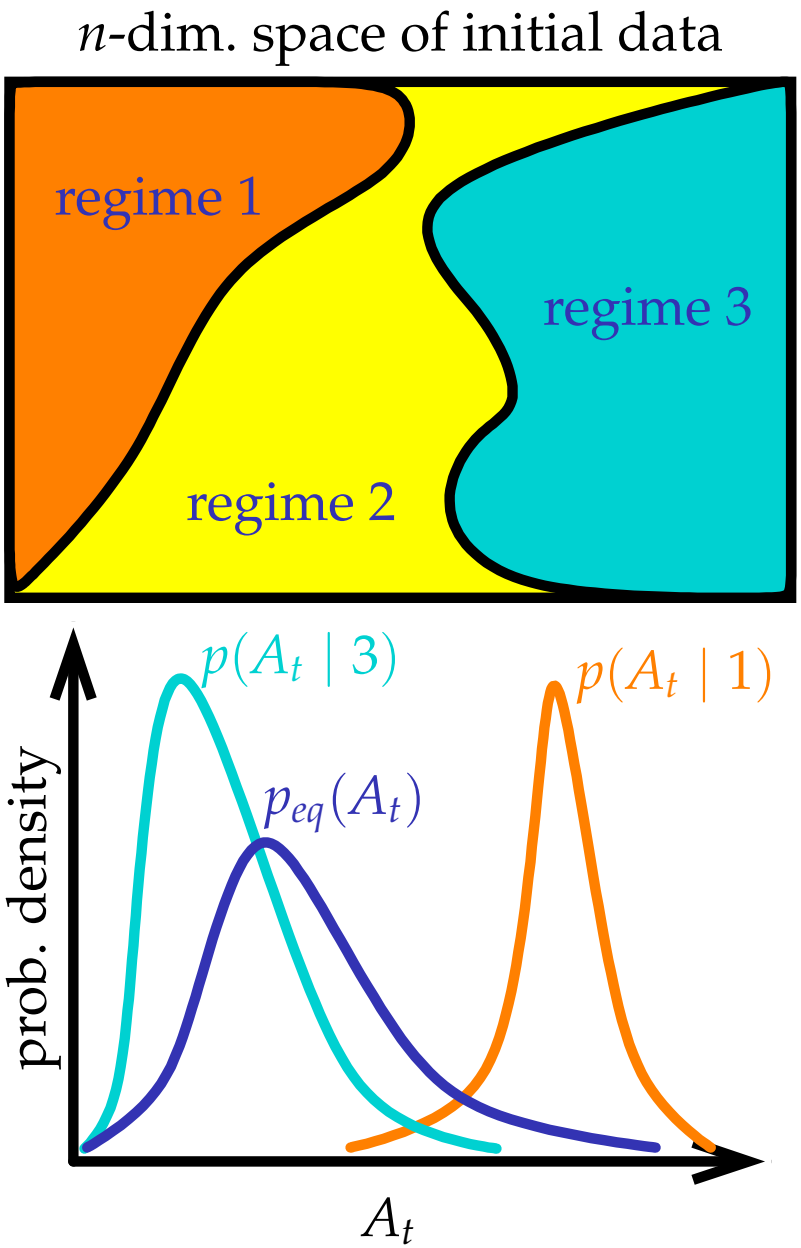




- Due to chaotic mixing, detailed initial data contribute negligible information for long-range forecasts.

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

Long-range, coarse-grained forecasts

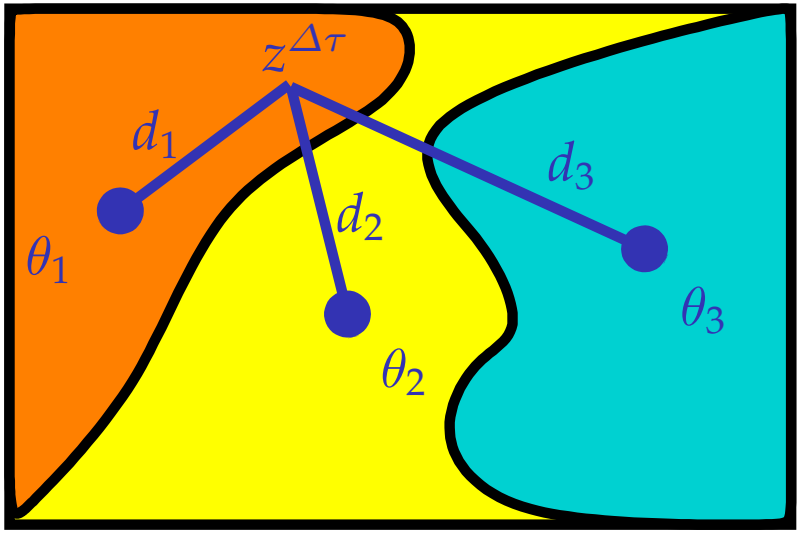


Strategy

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, θ_k.

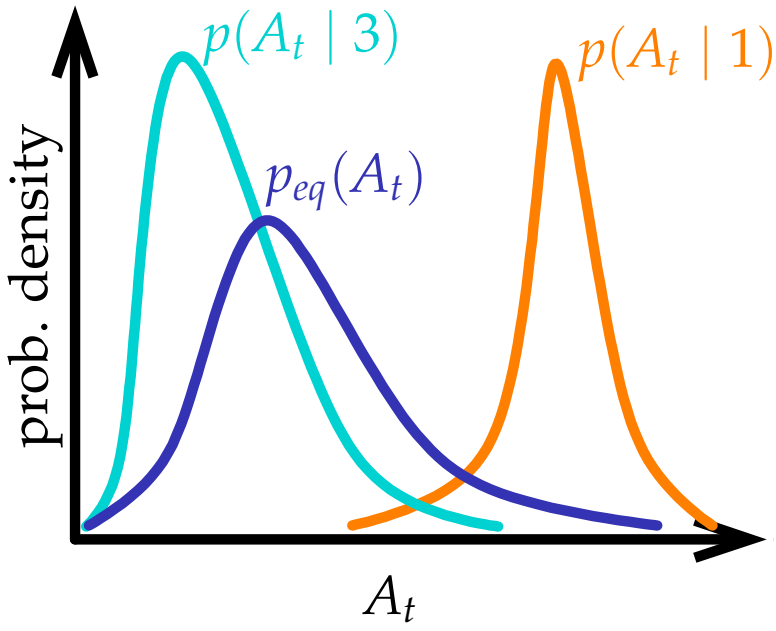
1. Collect observations $z(t)$ over a time window $\Delta\tau$ and compute the average,

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

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

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

The predictive information content in a partition



Predictive information given that the initial data lie in the k-th cluster:

$$\mathcal{D}_t^k = \mathcal{P}(p_t^k, p_{eq}), \quad p_t^k(A) = p(A_t \mid S = k).$$

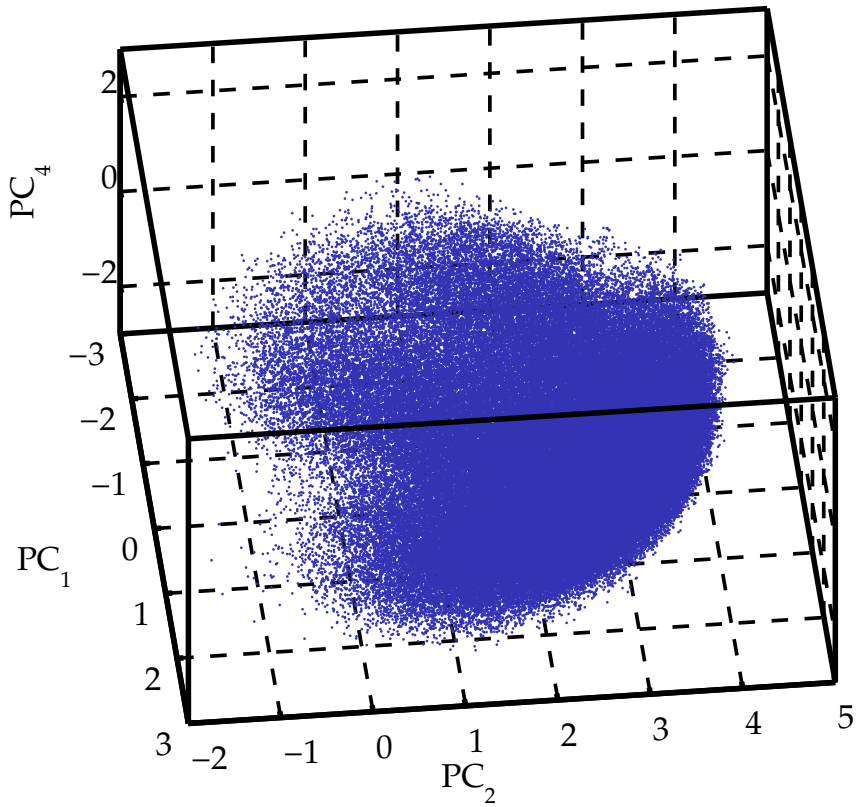
"Super-ensemble" measure of skill:

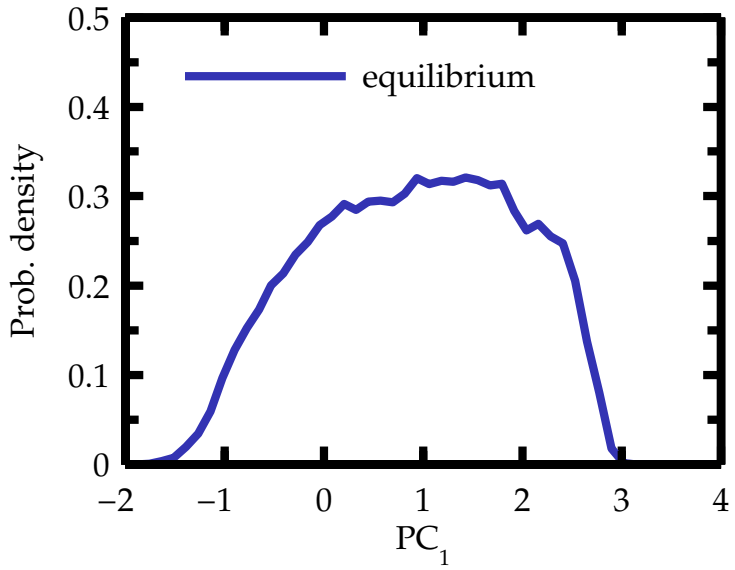
$$\mathcal{D}_t = \sum_{k=1}^{K} \pi_k D_t^k, \quad \pi_k = p(S = k).$$

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; namely, in the $t \to \infty$ limit.

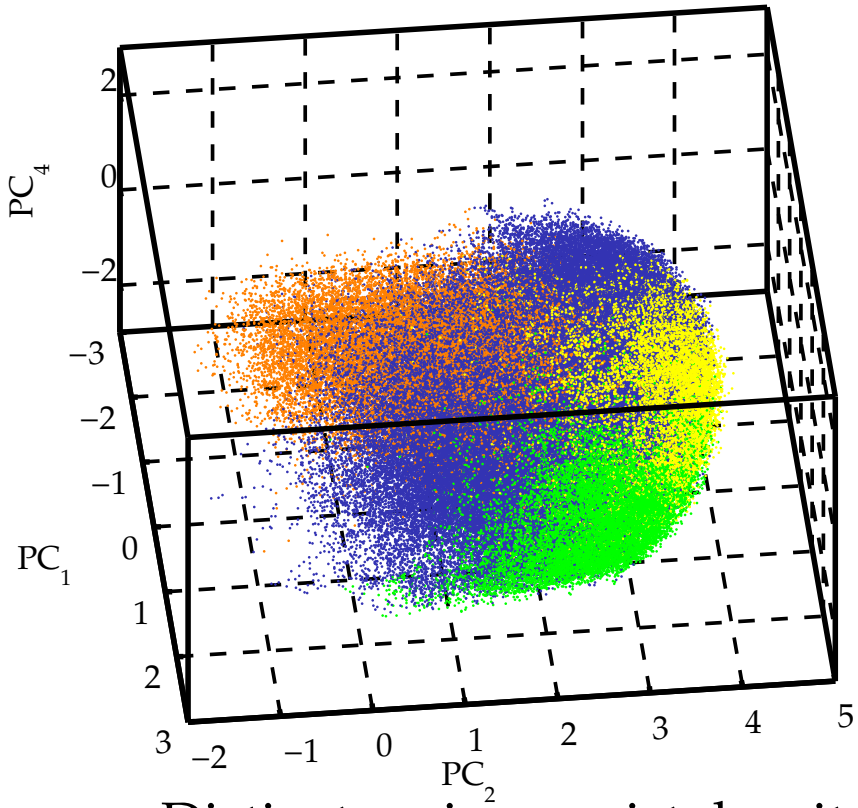
Coarse-grained partitions

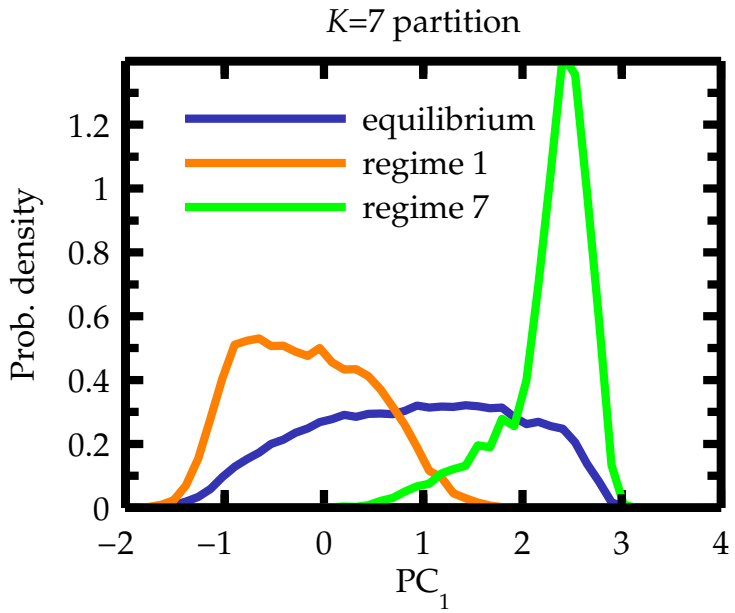




The point cloud of data in equilibrium.

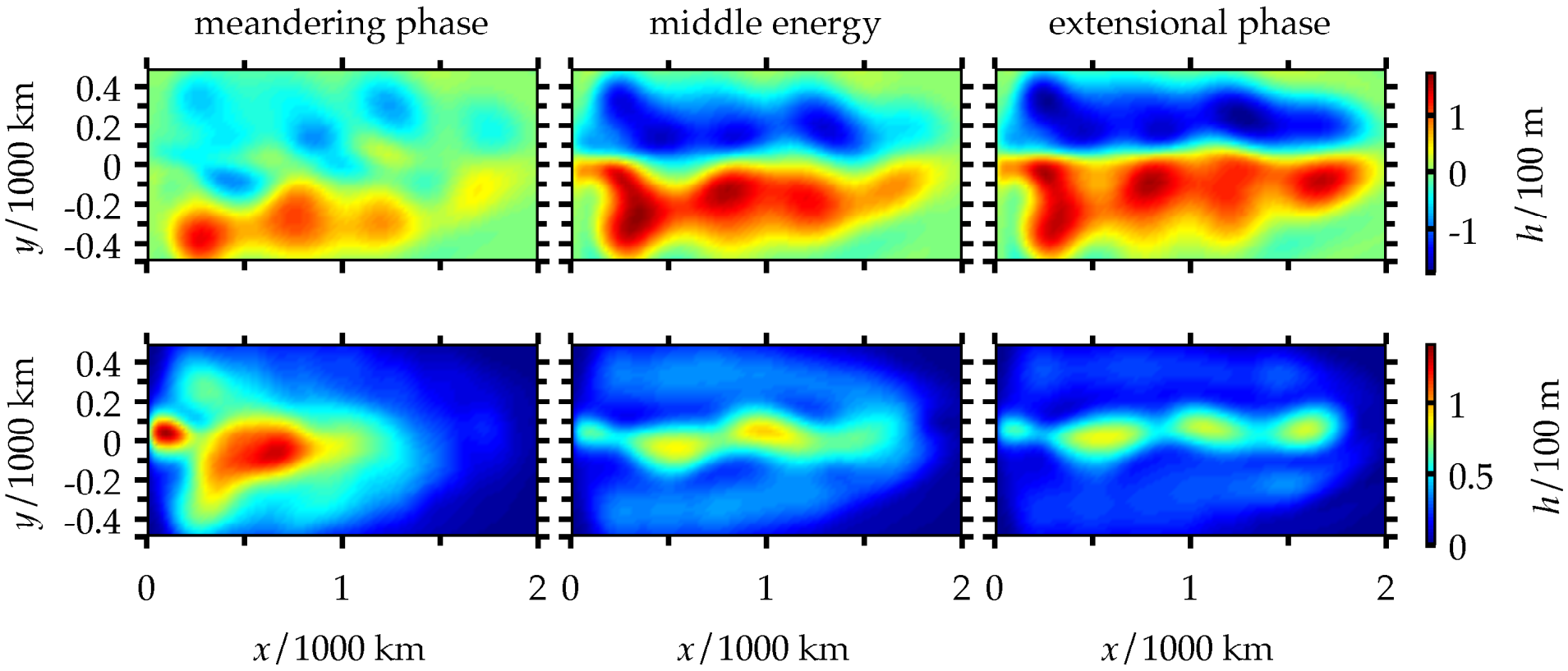
Coarse-grained partitions





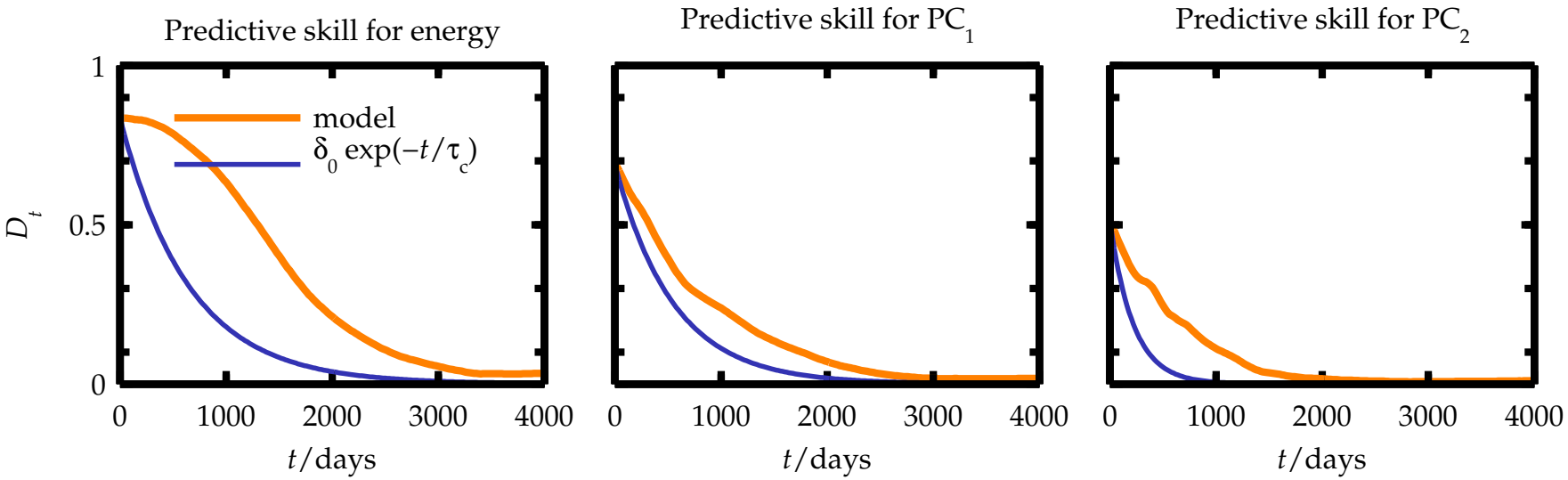
- Distinct regimes exist despite the lack of maxima in the equilibrium distributions of the PCs.
- Running-average smoothing of the training data and the initial conditions is crucial to reveal regimes.

Circulation regimes



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

Long-range forecasting skill



- Predictability decays more slowly than an exponential decay based on the decorrelation time τ_c.
- Energy is predictable up to ~ 7 years in advance.
- The leading PCs governing the large-scale structure of the flow are predictable up to ~ 5 years in advance.

Long-range forecasts with imperfect models

- Internal metric of "skill":

$$\mathcal{D}_t^M = \mathcal{P}(p^M(A_t \mid S), p_{eq}^M(A_t)).$$

- Equilibrium-consistency condition:

$$\mathcal{E}_{eq} = \mathcal{P}(p_{eq}(A_t), p_{eq}^M(A_t)) \ll 1. \qquad (*)$$

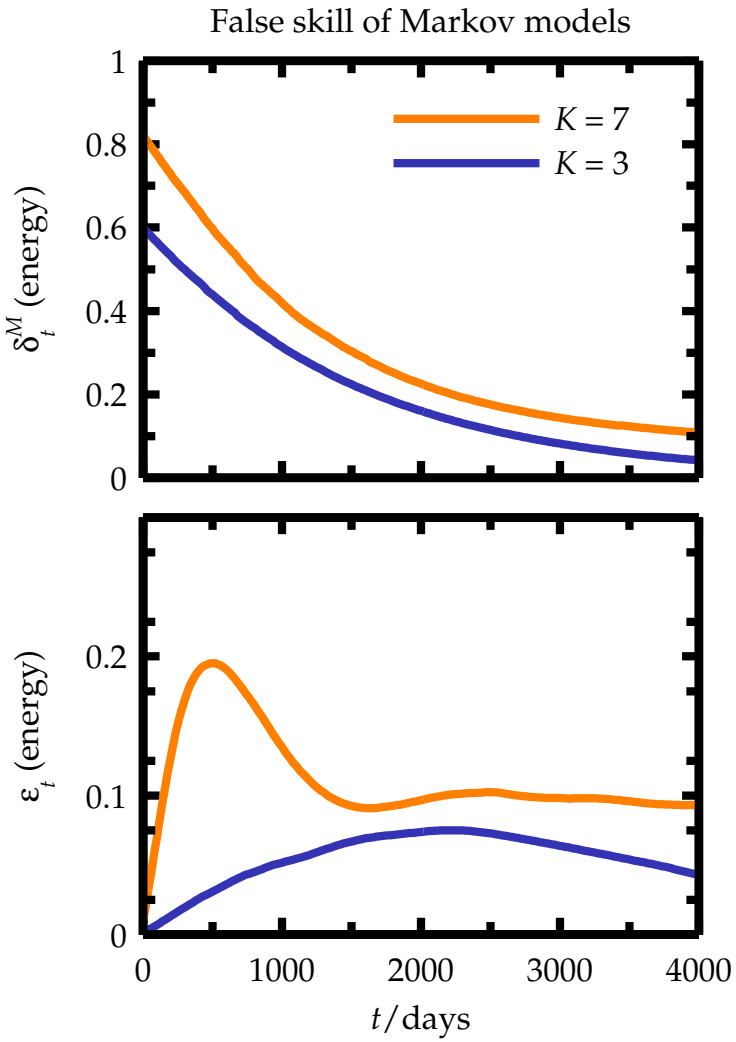
- Model error during relaxation to equilibrium:

$$\mathcal{E}_t = \mathcal{P}(p(A_t \mid S), p^M(A_t \mid S)).$$

Assessing models via relative entropy

- If $(*)$ is not met, $\mathcal{D}_t^M$ measures false skill at all times.
- At lead time t, $\mathcal{D}_t^M$ conveys genuine predictive skill only if $\mathcal{E}_t$ and $\mathcal{E}_{eq}$ are both small.

Markov models of regime transitions



Forecast PDFs in Markov models of regime transitions:

$$p^{Mk}(A_t \mid S) = \sum_{i=1}^{K} [\exp(\boldsymbol{L}t)]_{Si} p(A_0 \mid i).$$

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

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

Error in the Markov model relative to the perfect model:

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

- Equilibrium consistency is essential for long-range forecasts.
- Persistence of imperfect models is not synonymous with fidelity.

Conclusions

- Information-theoretic notion of predictability as the additional information beyond equilibrium about future values of observables gained by observing initial data S.
- Model error measured by the lack of information in forecast PDFs of imperfect models relative to a reference model (ideally, nature).
- For decadal initial-value 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 initial data (a major challenge in ensemble forecasts).
  - Clustering alleviates the curse of dimension.
- In this context, forecasts with imperfect should be assessed via:
  1. The fidelity of reproducing the equilibrium climate.
  2. Model error during relaxation to equilibrium.
  3. Amount of information provided beyond equilibrium.

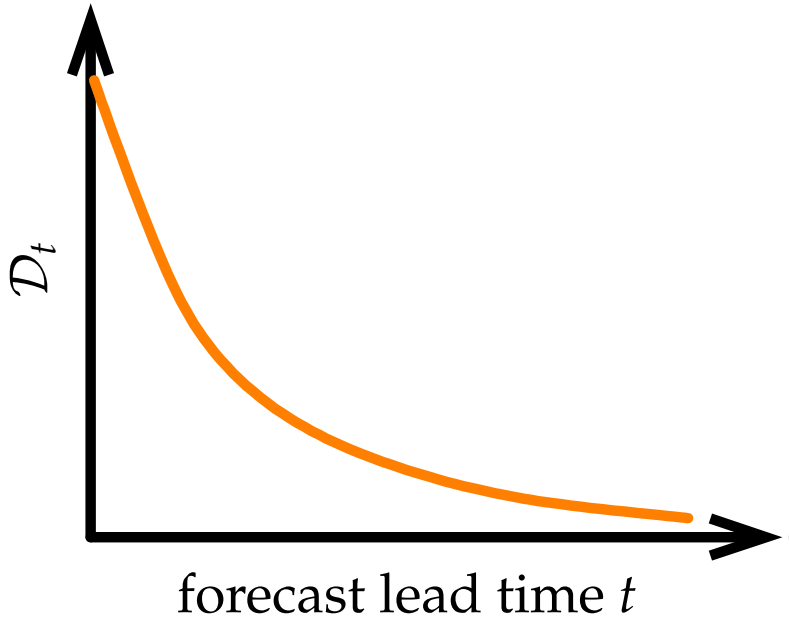
References

- 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.
- Kleeman (2010), Information theory and dynamical system predictability, *Isaac Newton Institute Preprint Series*, NI10063.

[this work]

- Giannakis & Majda (2011), Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model, *J. Climate* (to appear).
- 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, *J. Climate* (to appear).
- Giannakis, Majda & Horenko (2011), Information theory, model error, and predictive skill of stochastic models for complex nonlinear systems, submitted to *Physica D*.

Generalized second law



The relative entropy $\mathcal{D}_t$ is a non-increasing function of time if the following conditions are met:

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

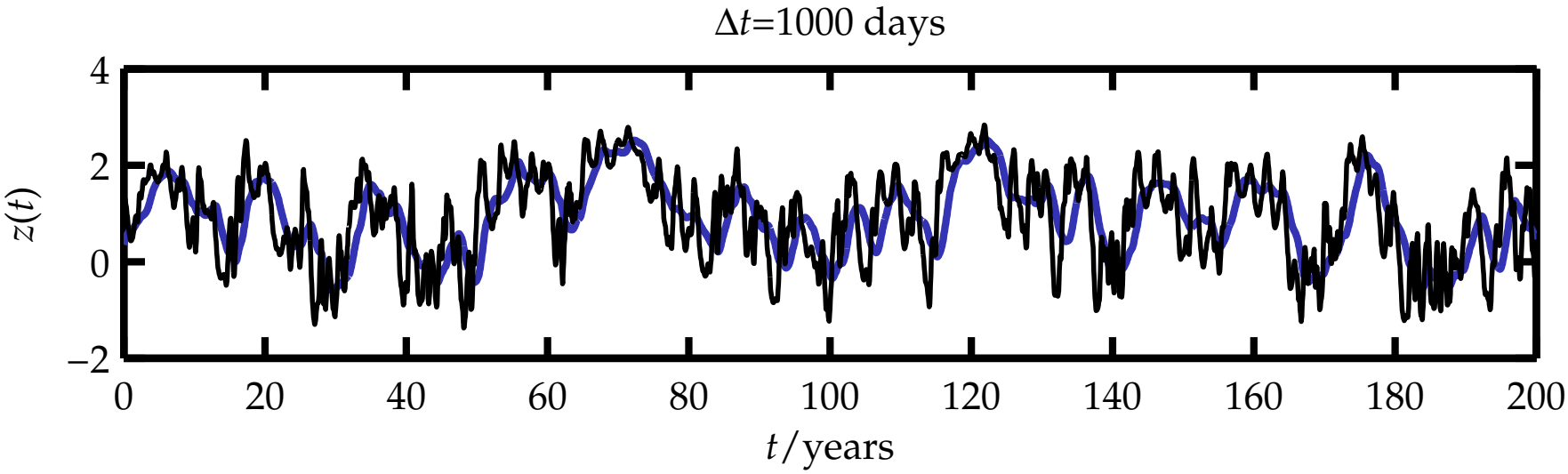
This means that the system must be *closed* and A_t has no *memory* or *hysteresis*.

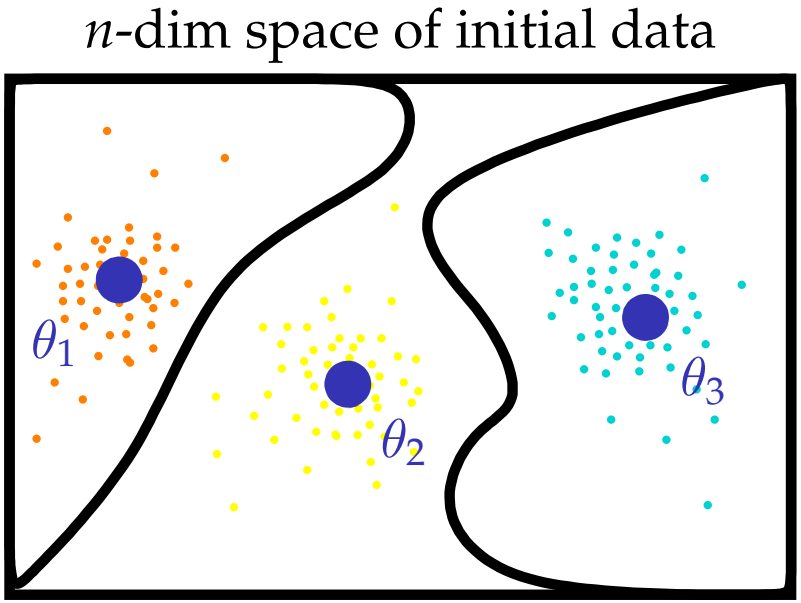
Setting

- 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),
$$A_t = A(x(t)).$$

Evaluating the cluster coordinates

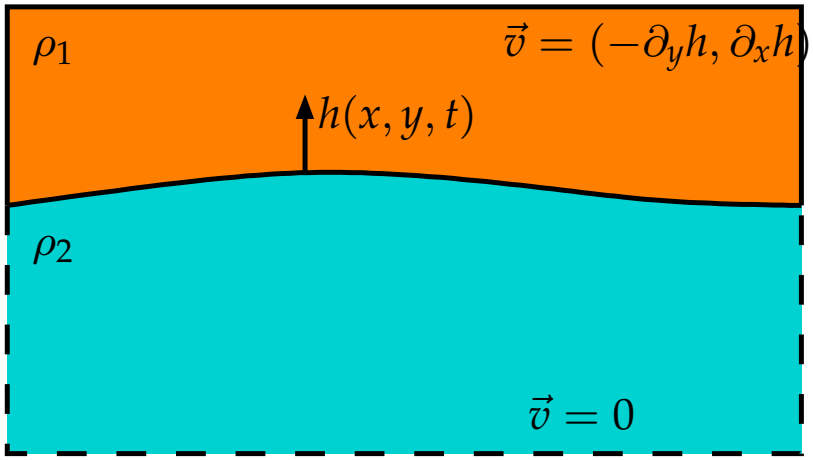
1. Collect a training time series $z(t)$, and take the running-average over a time window Δt.



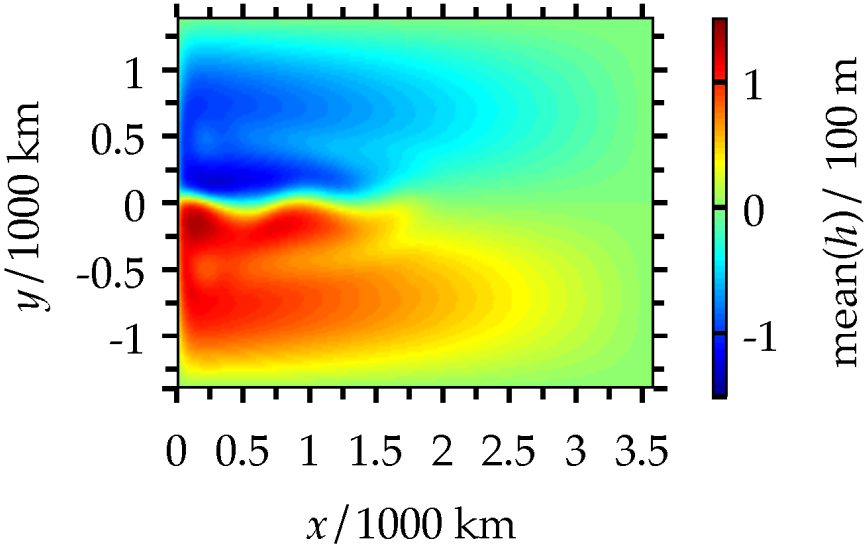


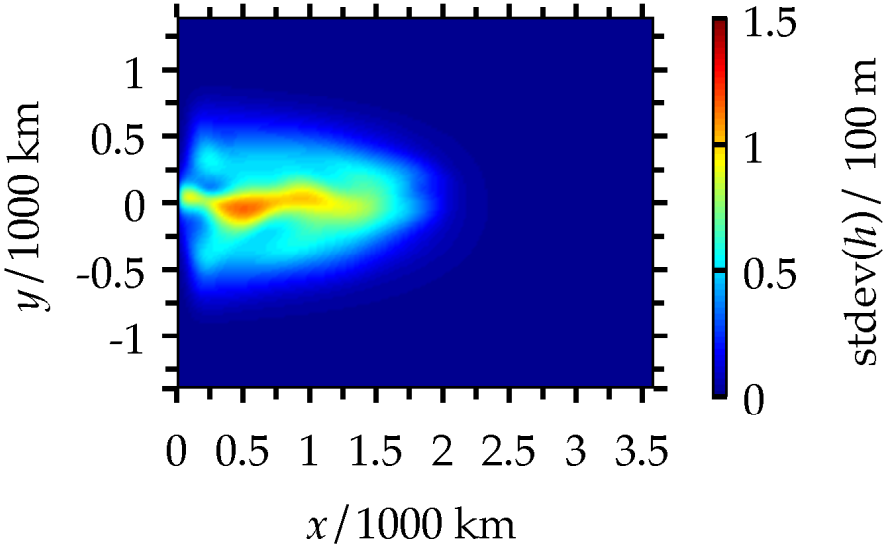
2. Determine θ_k by K-means clustering of $z^{\Delta t}(t)$.

The 1.5-layer model*



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





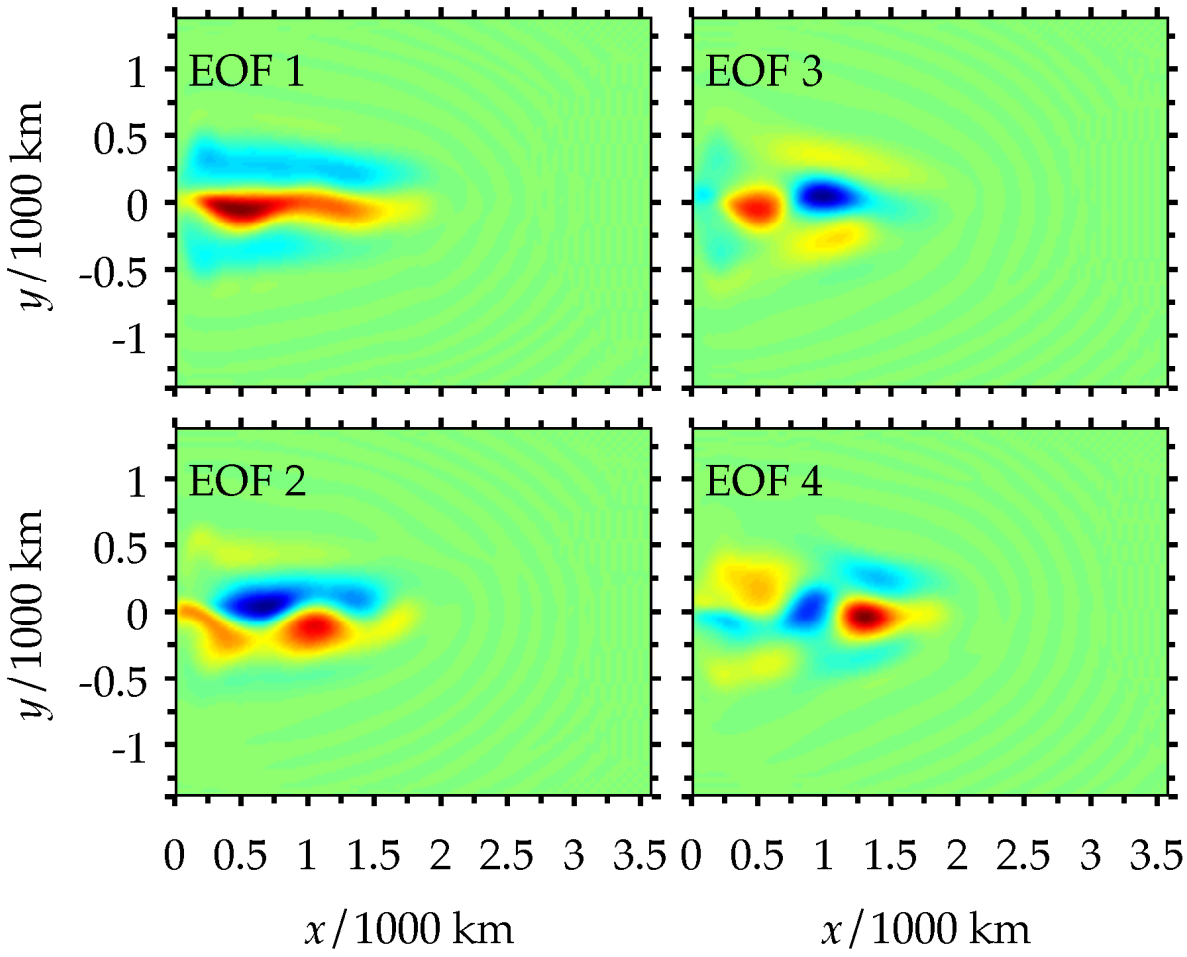
*McCalpin & Haidvogel (1996), *J. Phys. Oceanogr.*, 26, 739

Setup

Full dynamical system	$x(t) = h(\boldsymbol{r}, t)$
Incomplete measurements	$z(t) = (\mathrm{PC}_1(t), \ldots, \mathrm{PC}_{20}(t))$
Prediction observables	E (energy), PC_1, PC_2

- 20-dimensional measurement vector
- Scalar prediction observables

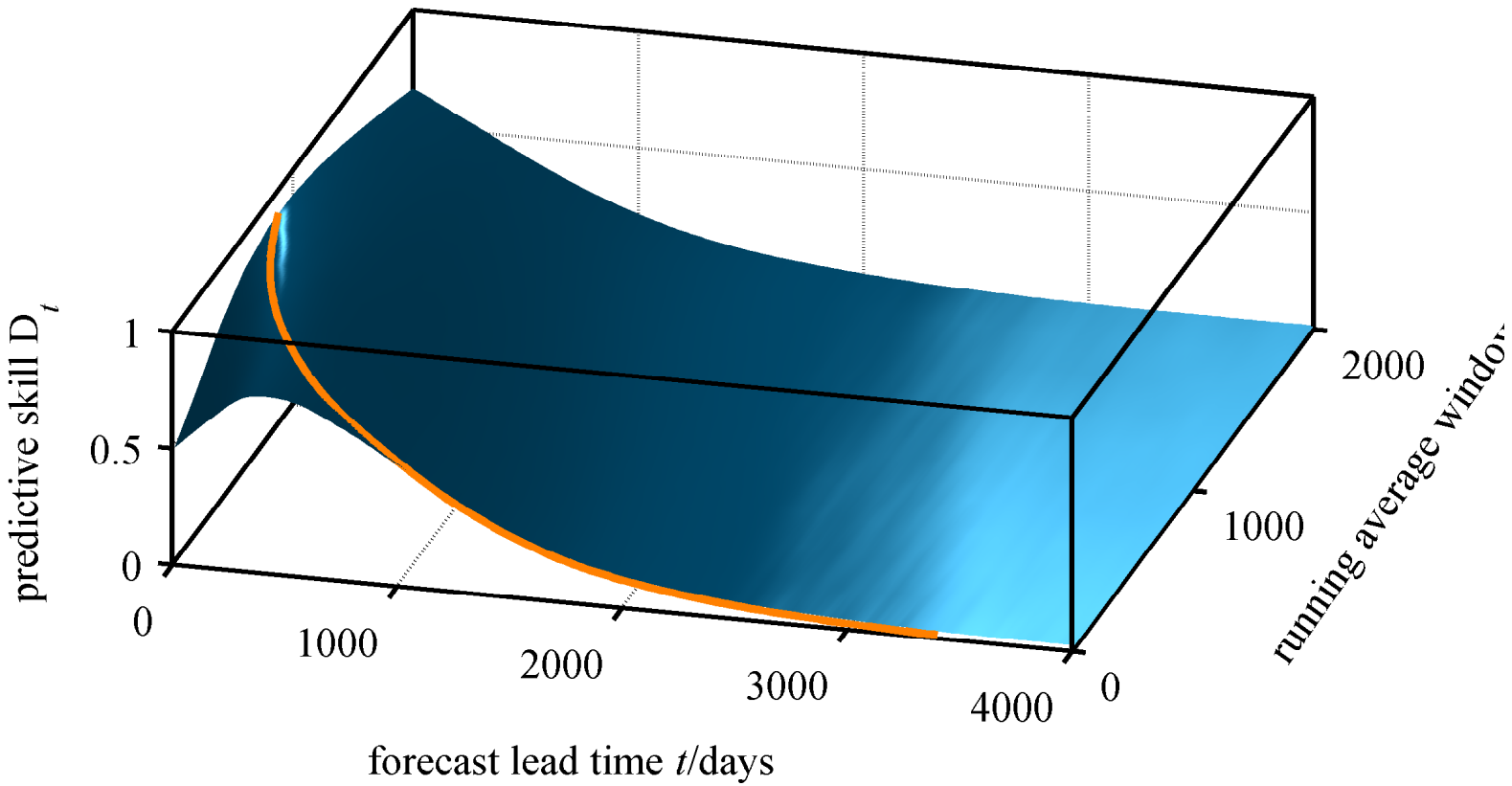
Empirical orthogonal functions (EOFs)



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

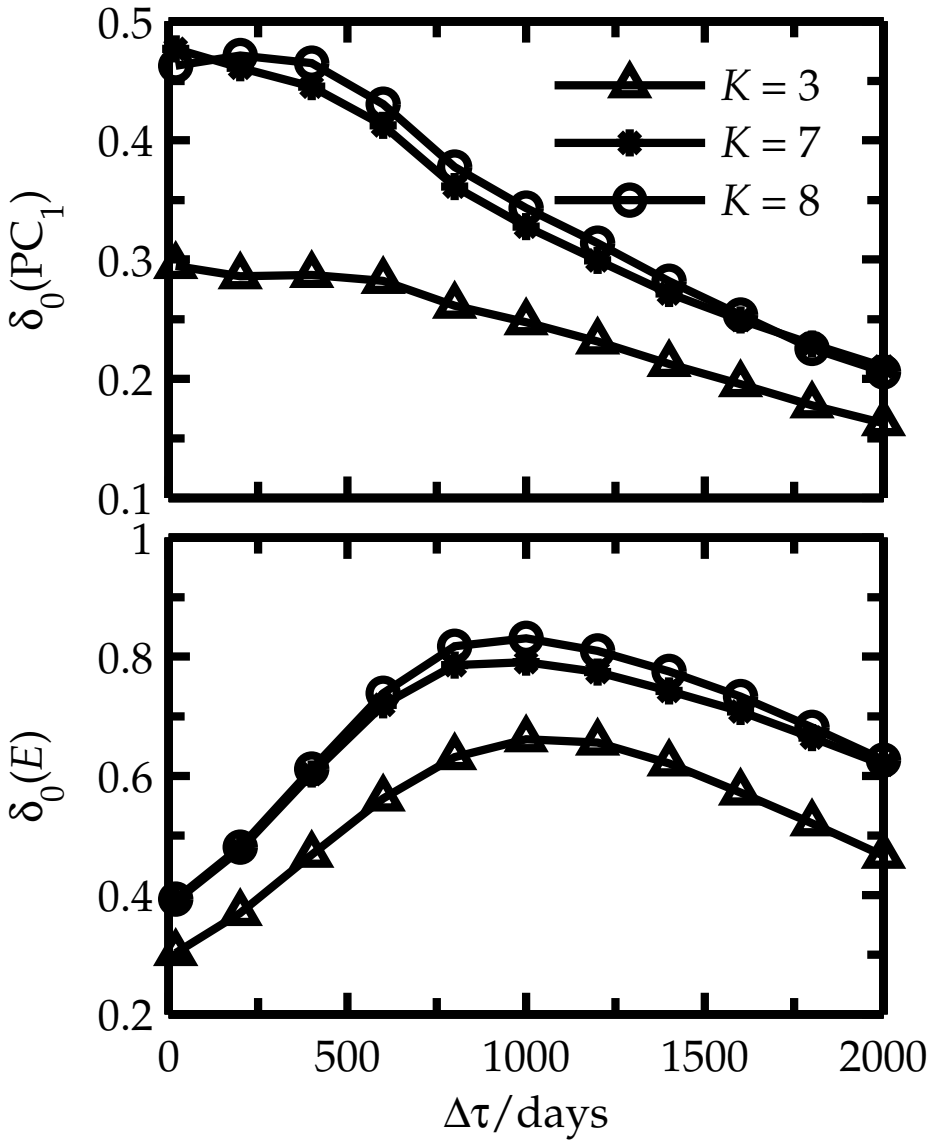
- PC_i is the principal component corresponding to EOF_i.
- For given n, this basis minimizes the norm of the residual.

Long-range forecasting skill



Predictive skill for energy. The optimal running-average window $\Delta\tau$ to preprocess the initial data maximizes $\mathcal{D}_t$ for the given prediction observable and forecast lead time.

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.