Information theory and statistical predictability Part I: Basic theory and simple models

Richard Kleeman
Courant Institute of Mathematical Sciences
New York

Some relevant references

T. M. Cover and J. A. Thomas. Elements of information theory. 2nd Edition Wiley, 2006.

R. Kleeman. Measuring dynamical prediction utility using relative entropy. *J. Atmos Sci,* **59**:2057-2072, 2002.

Information Theory Survey

Random Variables

A set of outcomes $\{x_i\}$ with an associated probability function $p(x_i)$ for each.

Information Theory

Originated from the study of the digital transmission of random variable outcomes (Shannon). The idea was to find the most efficient method/code. It turns out that the more uncertain a random variable the longer the (most efficient) code required. This can be precisely quantified and thus gives a universal measure of the uncertainty of a random variable. This functional is called the entropy.

Entropy Functional

If the set of outcomes is countable the entropy is given by

$$H(X) \equiv -\sum_i p(x_i) \log p(x_i)$$

This is maximized when p is uniform and is zero when the random variable is certain (exercise).

Conditional Entropy

Suppose we have another random variable Y associated with our system. If we were to know the outcome of this variable exactly then the uncertainty associated with X would be reduced. Write this as

$$H(X|Y = y)$$

In general Y is not certain and has a probability function $q(y_i)$ but we can still calculate the expected reduced uncertainty in X were we to know Y exactly. This is the conditional entropy

$$H(X|Y) = \sum_i q(y_i) H(X|Y = y_i)$$

Information Theory Survey

Relative Entropy

Sometimes we wish to know how much probability distributions differ from each other. The relative entropy is the most commonly used measure for this. It gives the informational inefficiency (in the Shannon sense) of assuming a random variable has a distribution $q(x_i)$ when if fact it has a distribution $p(x_i)$. For a countable set of outcomes it is given by

$$D(p||q) \equiv \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

This functional is always positive unless $p = q$ when it vanishes. It is not symmetric so is not a conventional distance function on distributions. Note that when $q(x_i)$ is uniform then the relative entropy becomes related to the ordinary entropy (exercise for the finite outcome case).

Mutual Information

It is often useful to know how related to each other two random variables X and Y are. In conventional statistics this is often measured by the correlation coefficient. Information theory offers a more universal measure of relationship. If the two variables were completely independent then their joint probability distribution would be

$$p(x_i, y_j) = p(x_i)q(y_j)$$

i.e. the probability of any particular outcome of X is unaffected by what the probability of any outcome of Y is.

We can use the relative entropy between the actual joint distribution and the independent

Information Theory Survey

one to measure independence. This is called mutual information

$$\begin{aligned} I(X;Y) &\equiv D(p(x_i,y_j)||p(x_i)q(y_j)) \\ &= \sum_{i,j} p(x_i,y_j) \log \frac{p(x_i,y_j)}{p(x_i)q(y_j)} \end{aligned}$$

It is easily shown the mutual information is related in an intuitively appealing way to ordinary entropy and conditional entropy. Indeed

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Thus the mutual information is the (expected) reduction in uncertainty in X caused by knowledge of Y and symmetrically the reduction in uncertainty of Y caused by knowledge of X.

Errors in Predictions

- The particular physical model used may be flawed. Commonly called model error. Difficult to analyze since such errors have little commonality.
- In a closed dynamical system there may be initial condition errors. Such errors are unavoidable in any practical prediction since observing systems are always incomplete and subject to some error.

In this talk we focus on errors of the second kind. This is often called a perfect model prediction scenario. We shall also focus on the dynamical evolution of such errors rather than on their specific form in the initial conditions since the latter is obviously observing system dependent. Errors of the first kind are obviously important but in a sense are issues of better future modelling rather than fundamental limitations. The presence of uncertainty implies that formally variables should be considered random variables each with an associated probability distribution.

A Bayesian Approach to Utility

Suppose we have a vector of random variables **X** with vector outcomes **x** associated with a particular dynamical system. Further suppose that such a system prior to the present has been observed for a significant period and the probability distribution q(**x**) deduced. Now suppose that further observational information concerning the system is received. In general this will result in a modified distribution p(**x**). In the Bayesian perspective q(**x**) is referred to as the prior distribution whereas p(**x**) is called the posterior distribution. Bayesian learning theory measures the utility of the additional observations using the relative entropy of the two distributions.

The process of statistical prediction in a dynamical system fits this model of learning: If no reference is made to time and the associated state of the system (i.e. the initial conditions) then the natural choice for a prior distribution is provided by the equilibrium or climatological distribution. Suppose now that the initial state is carefully observed and the state projected forward in time using a dynamical model. This process will result in a modified distribution for the random state variables at some future time. Such a distribution is commonly called the prediction distribution. The discrepancy or relative entropy between the prediction and equilibrium distributions then is a measure of the utility of the prediction process.

Properties of Relative Entropy

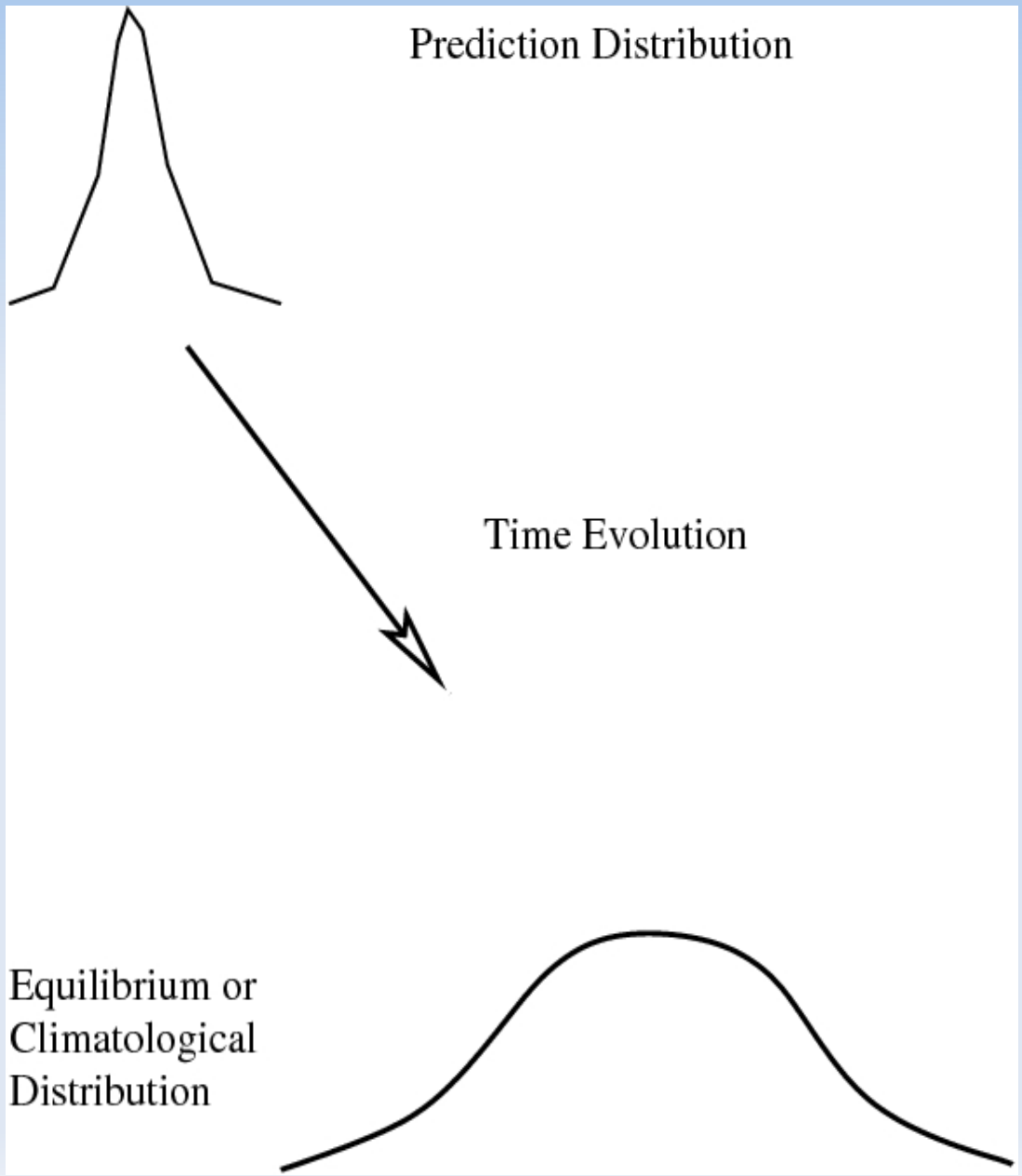
$$D(p\|q) \equiv \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$$

- Relative entropy is non-negative and zero only when the distributions are effectively identical. It is asymetric which is perhaps not surprising since so is the (ideal) learning process.
- In a closed system relative entropy does not increase with time and usually decreases. The Earth overall is approximately a closed system in the sense meant here.
- If one applies a general non-linear non-singular transformation to the state space then relative entropy is left invariant. Thus it is a 'universal' measure of utility. Very few skill/utility measures have such a property.

Given the above attractive mathematical properties we use relative entropy as a measure of prediction utility.

Utility as disequilibrium

- In conventional stochastic differential equation theory, convergence of distributions to equilibrium is measured by the relative entropy (called the Lyuponov functional) since it satisfies the temporal monotonicity property above.
- In practical dynamical systems asymptotic convergence to a unique equilibrium distribution is common. Thus it is illuminating to use relative entropy as a measure of both predictability as well as of statistical disequilibrium.
- Notice that upon convergence of prediction distributions (as measured by relative entropy), initial condition information becomes irrelevant.



Gaussian Distributions

Many practical applications have approximately Gaussian pdfs so it is of interest to use the analytical formula for Gaussian relative entropy.

The general multivariate form of the Gaussian distribution for a random n dimensional vector $\mathbf{x}$ can be written as

$$p(\mathbf{x}) = N \exp\left[-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^t \boldsymbol{\sigma}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right]$$

$$N = [(2\pi)^n \det(\boldsymbol{\sigma})]^{-1/2}$$

where $\boldsymbol{\sigma}$ is the $n \times n$ covariance matrix for the random variables of interest.

The relative entropy of two Gaussian distributions is given by

$$D(p||q) = Dispersion + Signal$$

$$Dispersion = \frac{1}{2}\left[\log\left(\frac{\det(\boldsymbol{\sigma}_q)}{\det(\boldsymbol{\sigma}_p)}\right) + tr\left(\boldsymbol{\sigma}_p\boldsymbol{\sigma}_q^{-1}\right) - n\right]$$

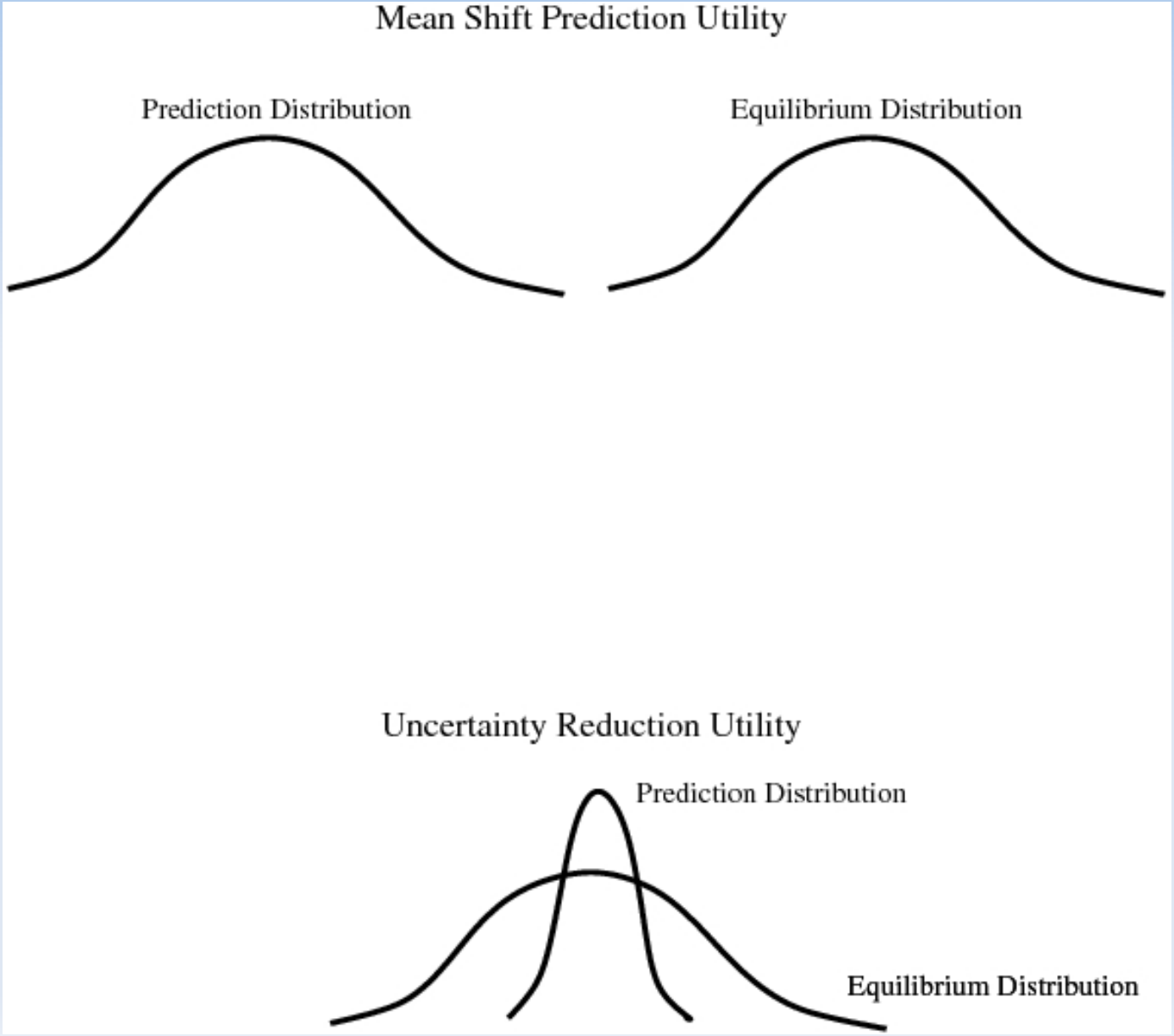
$$Signal = \frac{1}{2}(\overline{\mathbf{x}}_p - \overline{x}_q)^t \boldsymbol{\sigma}_q^{-1}(\overline{\mathbf{x}}_p - \overline{x}_q)$$

One dimensional version is more revealing intuitively

$$Dispersion = \frac{1}{2}\left[\log\left(\frac{\boldsymbol{\sigma}_q}{\boldsymbol{\sigma}_p}\right) + \boldsymbol{\sigma}_p\boldsymbol{\sigma}_q^{-1} - 1\right]$$

$$Signal = \frac{1}{2}\frac{(\overline{\mathbf{x}}_p - \overline{x}_q)^2}{\boldsymbol{\sigma_q}}$$

Meaning of Dispersion and Signal

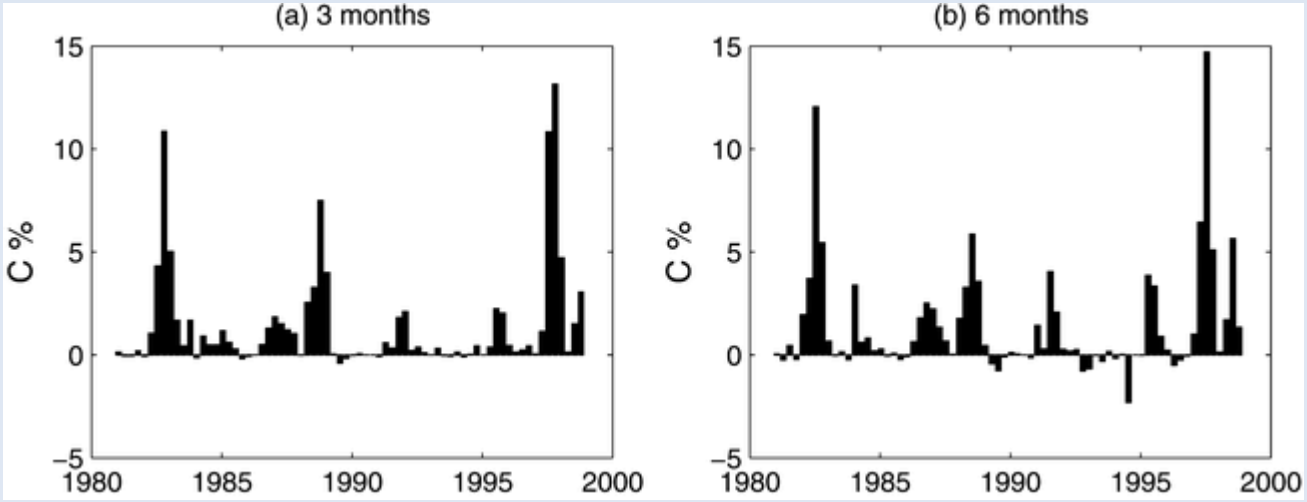


Why are models skillful?

Coupled models are able to predict ENSO with some skill for about 9-12 months.

Where does this skill come from?

The answer depends on the skill measure used obviously however for the usual anomaly correlation statistic used for NWP and ENSO prediction the answer is rather interesting for all the ENSO coupled models I have checked.....



If you break the anomaly correlation into its contributing pieces timewise you find nearly all the skill comes from major events. The larger the event the greater its contribution to skill.....

This is fascinating because it suggests that useful forecasts are not common but may be very useful indeed on certain occasions. We turn to theory to develop understanding.

Connection to usual skill measures

The relative entropy utility measure is defined on one particular set of ensemble predictions all starting from close to the same initial conditions. Typical skill measures are defined with respect to a series of predictions from different initial conditions. The former measure is a perfect model one while the latter are not.

We can make a (rough) connection between the two types of measures by considering the optimal case for prediction i.e. when the model is perfect and compute the common skill measures. This then represents an upper limit on skill given the model is a good one physically.

The two most common skill measures are RMS error and anomaly correlation. The theoretical upper limits of these are easily shown to be:

$$RMSE = \overline{\sigma}_p$$

$$AC = \frac{\overline{(\overline{x}_p - \overline{x}_q)^2}}{\overline{\sigma}_p + \overline{(\overline{x}_p - \overline{x}_q)^2}}$$

Where the upper overbar means average over all initial conditions. We see that AC (in an average and Gaussian sense) incorporates part of the signal term while RMSE is only related to dispersion. Due to the above formula AC can be viewed intuitively as a signal to noise metric which is consistent with the relationship we are attempting to draw with the signal part of relative entropy.

A simple stochastic oscillator

In the first lecture the stochastic oscillator was put forward as a believable model for ENSO. Let us therefore examine the simplest possible model for this and analyze its predictability characteristics.

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t = \begin{pmatrix} 0 & 1 \\ \beta & \gamma \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} 0 \\ F \end{pmatrix}$$

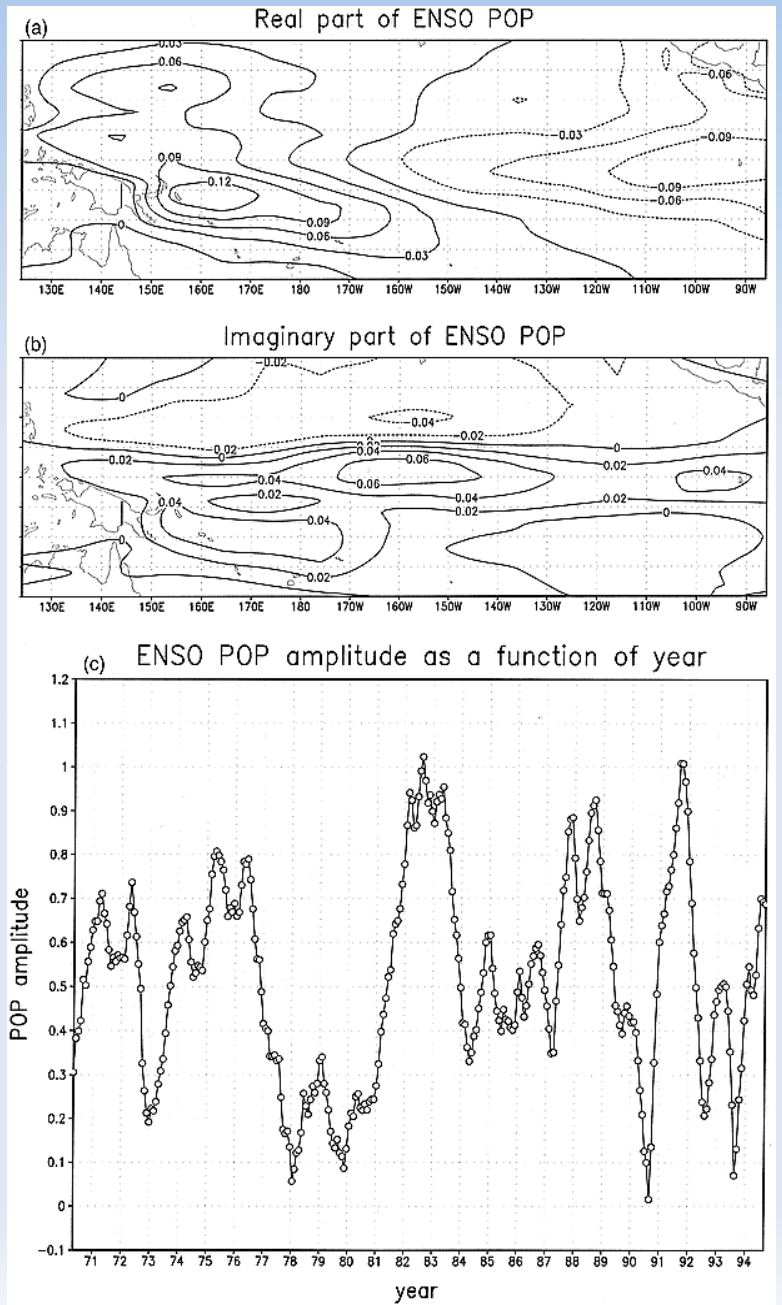
$$\gamma = -\frac{2}{\tau} \qquad \beta = -\left(\frac{4\pi^2}{T^2} + \frac{1}{\tau^2}\right)$$

The two components here can be interpreted as the coefficients of the first two EOFs of ENSO upper ocean heat content.

F is a white noise forcing and τ is the damping time while *T* is the period of the damped oscillator.

Noise represents fast modes. Simplest strategy for initial conditions is to assume very small errors in the fast modes which can be approximately represented as slow mode deterministic initial conditions. We choose these slow mode initial conditions at random from a realization of the equilibrium distribution since this amounts to a representative selection of initial conditions.

A simple stochastic oscillator



To be more concrete about our simple stochastic oscillator, the first two EOFs of upper layer heat content are shown at the left. Because they are correlated if lagged in time they can be grouped together as a Principal Oscillation Pattern (POP).

This time lag correlation or POP behaviour is completely consistent with our simple model of ENSO. One of the EOFs is the peak warm or cold event behaviour while the other is the classical Wyrtki build up of heat content which tends to precede events.

The amplitude of the POP which is central to predictability (see next slide) shows substantial variation with time as the bottom panel illustrates. Notice it was very small in the late 1970s and early 1990s.

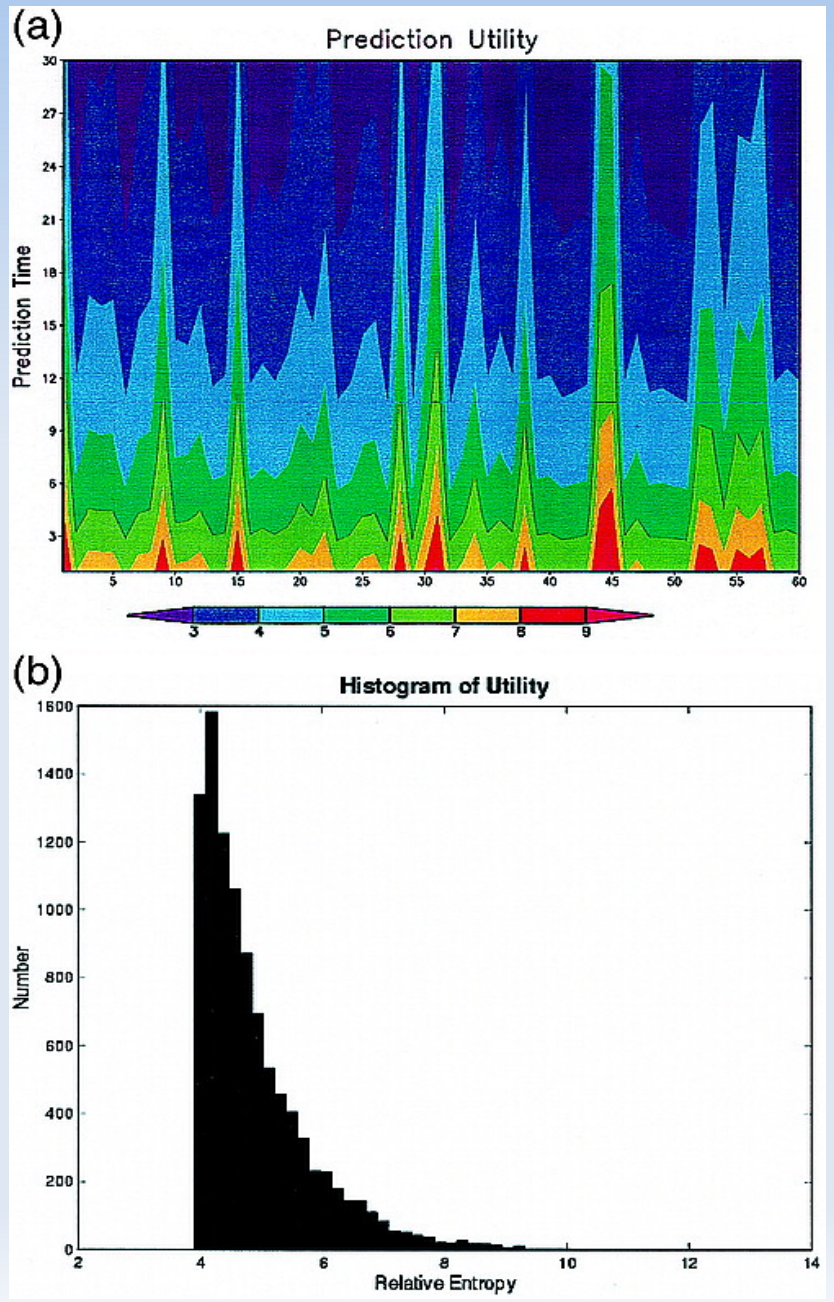
A simple stochastic oscillator

If the initial conditions are chosen to be deterministic then one can show the following important facts about solutions to this model

- The pdfs are (bivariate) Gaussian.
- The time evolution of the covariance matrix does not depend on the initial condition chosen.
- The (unique) equilibrium covariance matrix is diagonal i.e. the two model variables are uncorrelated at very long times.

The implication of these facts is that only the signal controls variations in utility with initial conditions. Secondly the signal component of utility is proportional to a (rescaled) square of the amplitude of the anomaly of the two components.

This centrally important amplitude or signal can vary with time just for random reasons i.e. it executes a random walk.



Signal based predictability