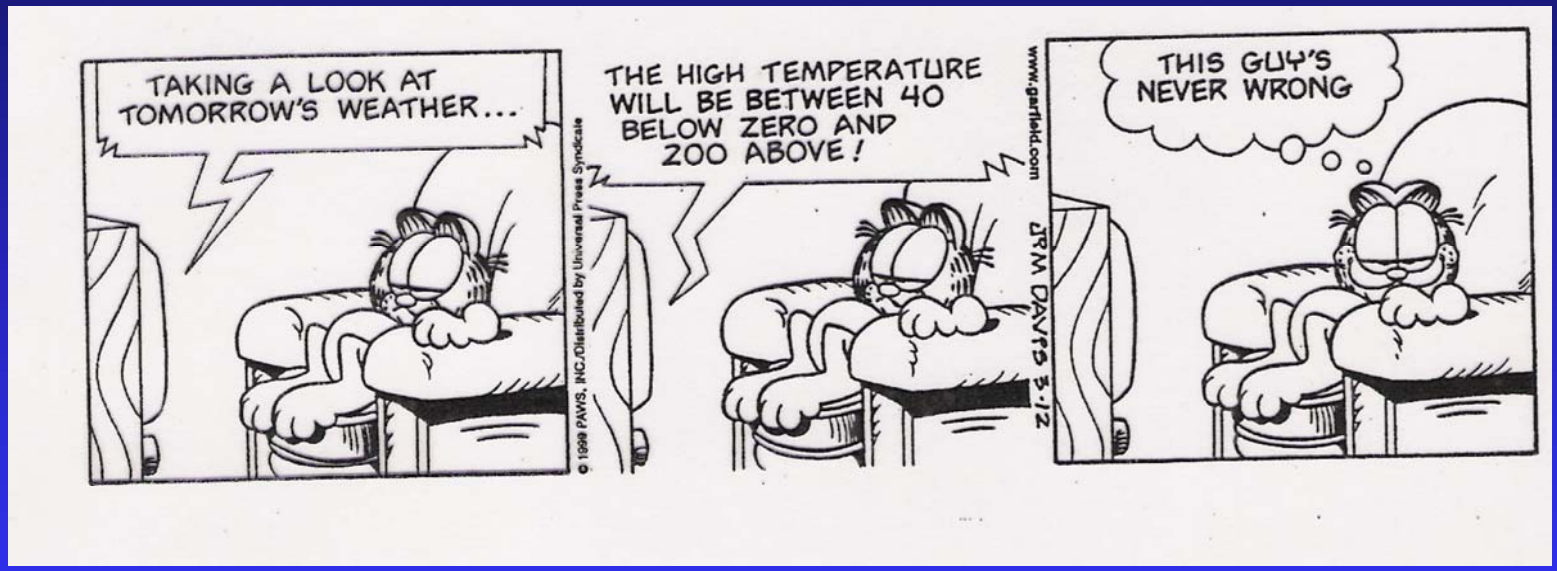


2 Feb 2006

Disentangling Uncertainty and Error: On the predictability of nonlinear systems.

Modern weather forecasters have taken the notion of uncertainty in the initial condition to heart. As a result, American, European and Canadian forecast centres run Monte Carlo style operational forecasts: every day an ensemble of different initial conditions are evolved under state-of-the-art weather models to obtain a forecast with quantitative forecast-uncertainty information. There are many opportunities for developing a more coherent statistical approach to this problem, ranging from aspects of ensemble design and forecast interpretation to socio-economic decision support, and even improved feedback to the modellers the enhance model improvement. Starting with a basic introduction to numerical weather prediction (assuming no meteorological background), this talk will survey a number of strengths, weaknesses, and needs of operational weather forecasting, and the implications these hold for both scientific impacts in and socio-economic impacts of meteorology and Earth Systems science.

2 Feb 2006

2 Feb 2006

Disentangling Uncertainty and Error: On the predictability of nonlinear systems

Leonard Smith

Centre for the Analysis of Time Series

London School of Economics

Pembroke College, Oxford

Mark Roulston, Devin Kilminster, Kevin Judd,
Liam Clarke, Jochen Broecker

lsecats.org

(Discussion Forums Now on-line)

Overview

- First we'll look at uncertainty
- Then an probability
- Then at Error and perhaps a few ways forward

This afternoon I'll talk about sources of value.

But why worry about this difference?

Because we want to improve forecasts!

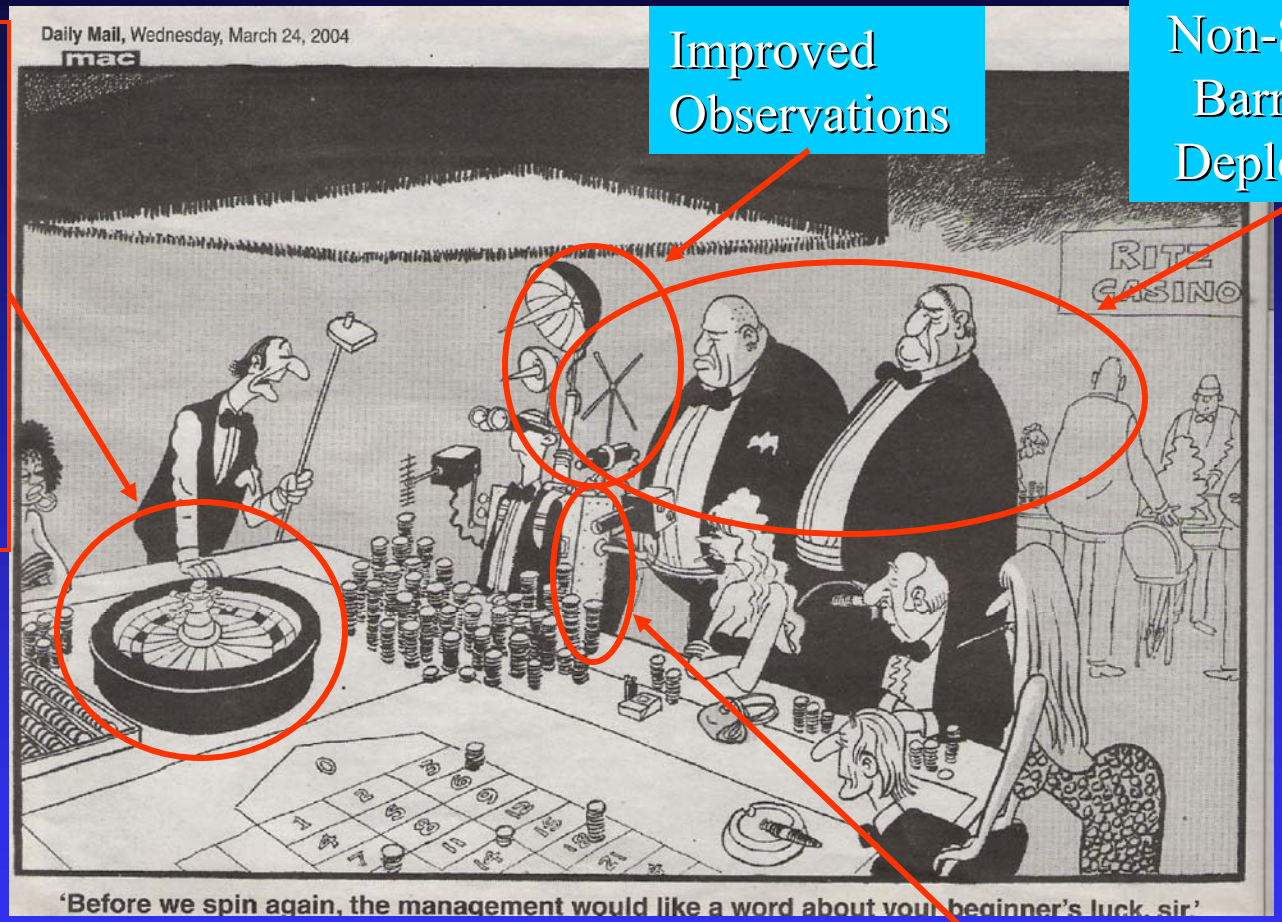
Alt: If we had enough data, could we predict more profitably?

2 Feb 2006

2) Resource allocation (Identifying Weakest Links)

Theory of Stochastic Processes

Classical Dynamics



Improved Observations

Non-Science Barriers to Deployment

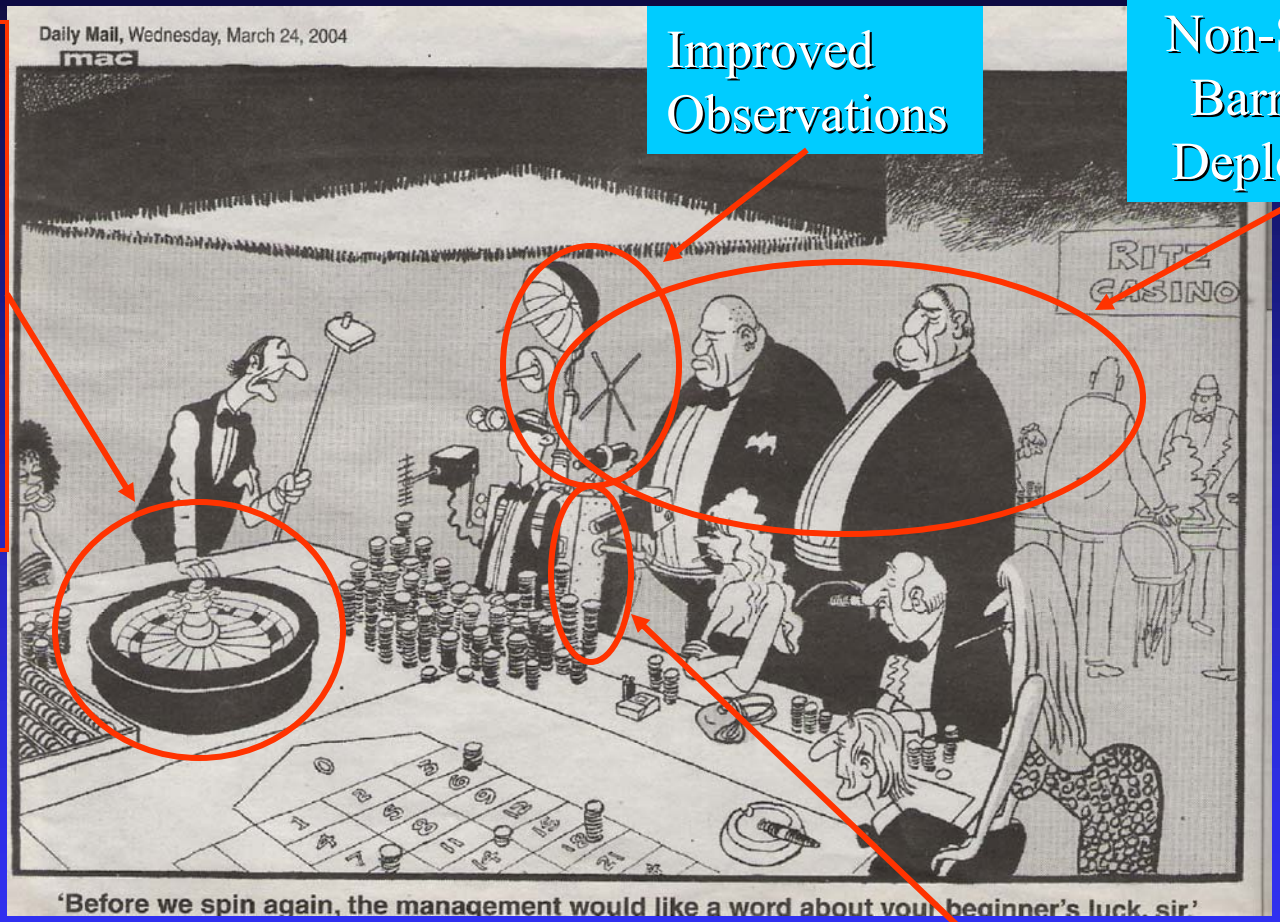
Bigger Computer

(In this particular case, obs were more valuable than theory)

Resource allocation
(Identifying Weakest Links)

Theory of Stochastic Processes

Classical Dynamics



Improved Observations

Non-Science Barriers to Deployment

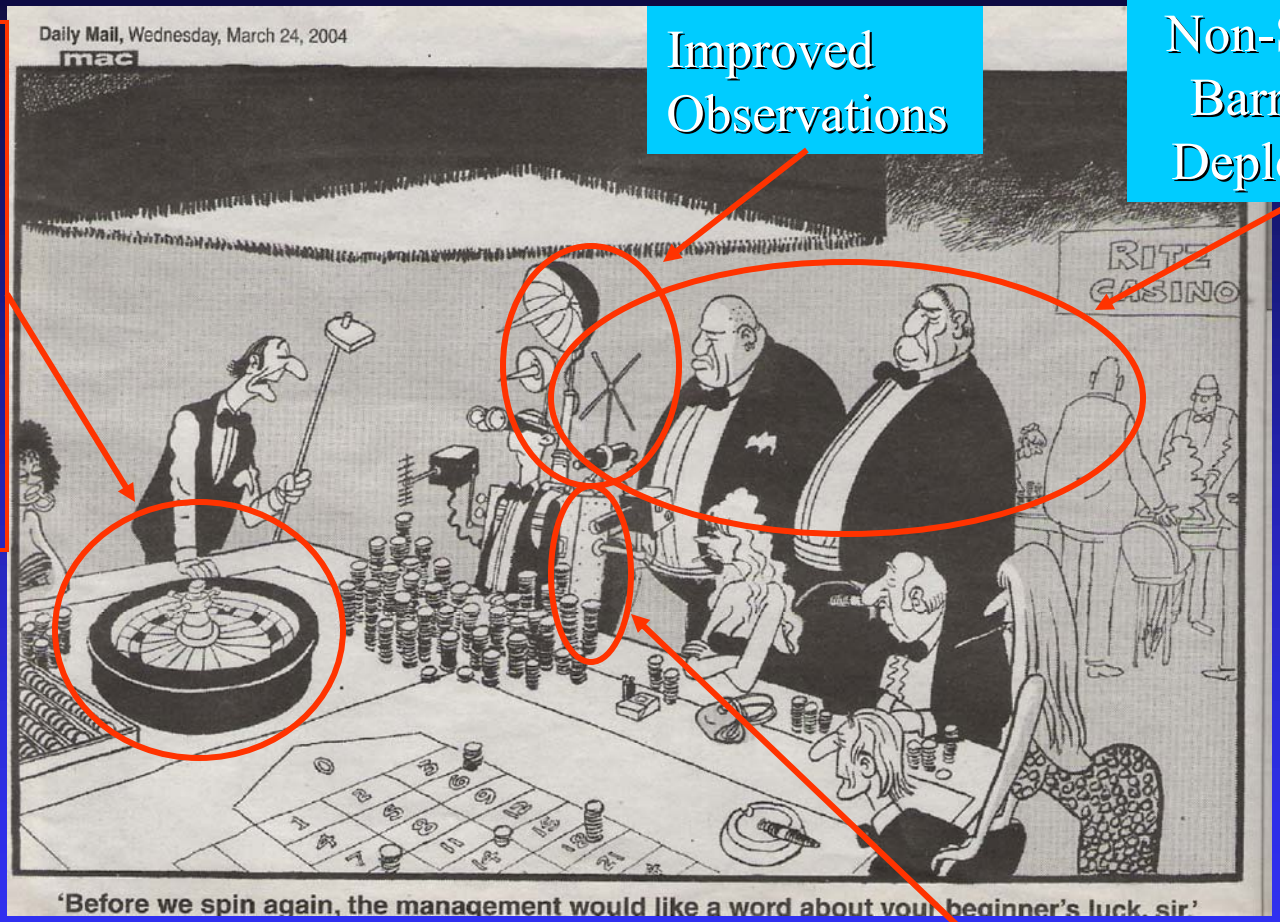
Bigger Computer

Before we improve "predictability", we need to measure it!

Resource allocation
(Identifying Weakest Links)

Theory of Stochastic Processes

Classical Dynamics



Improved Observations

Non-Science Barriers to Deployment

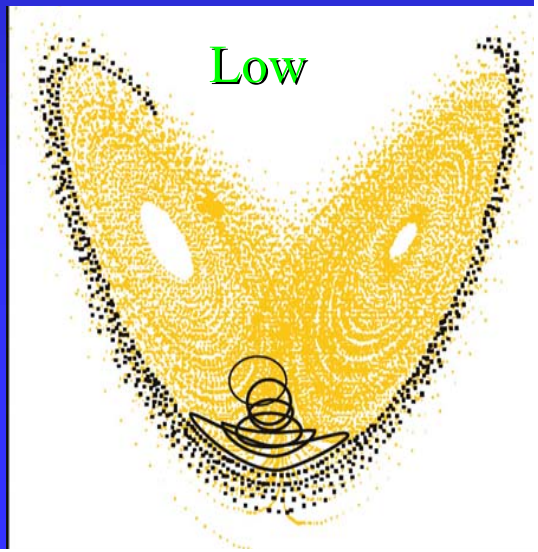
Bigger Computer

(In this particular case, obs are more valuable than theory)

What does *Predictability* look like?

But is the goal a vague indication of uncertainty? Or an accountable Monte Carlo style PDF?

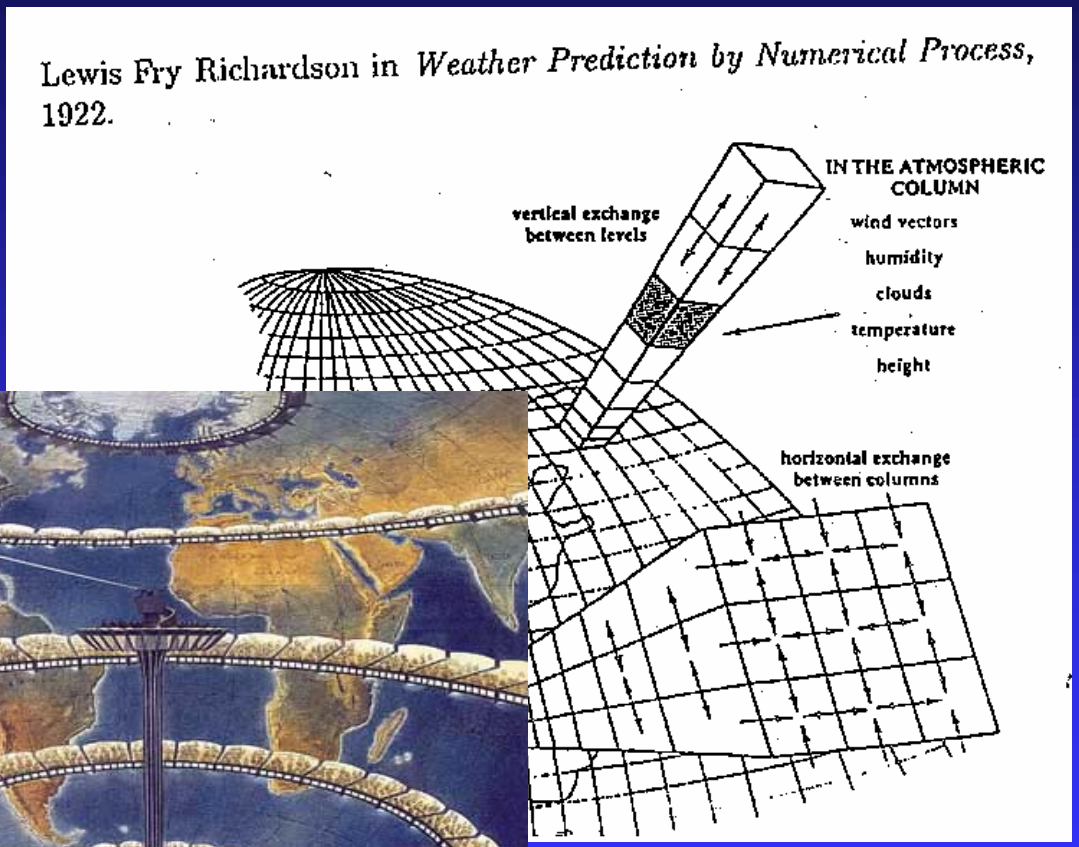
Pictures from Tim Palmer

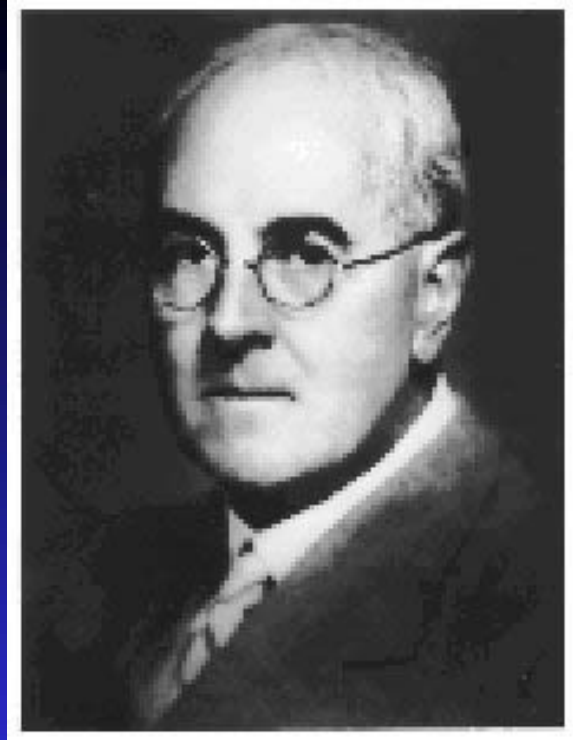


We would like to quantify day to day variations in predictability…

2 Feb 2006

L F Richardson made the first modern numerical weather forecast during WWI.





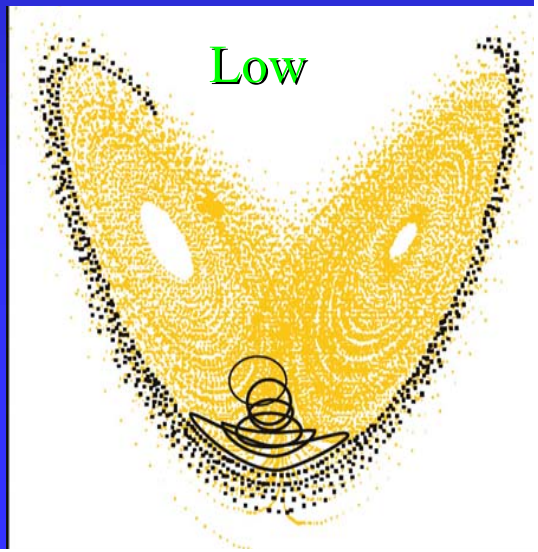
The original forecast was once lost during a retreat, only to be found weeks later under a pile of coal!

Today's NWP models follow the same basic plan.

2 Feb 2006

What does *Predictability* look like?

Pictures from Tim Palmer



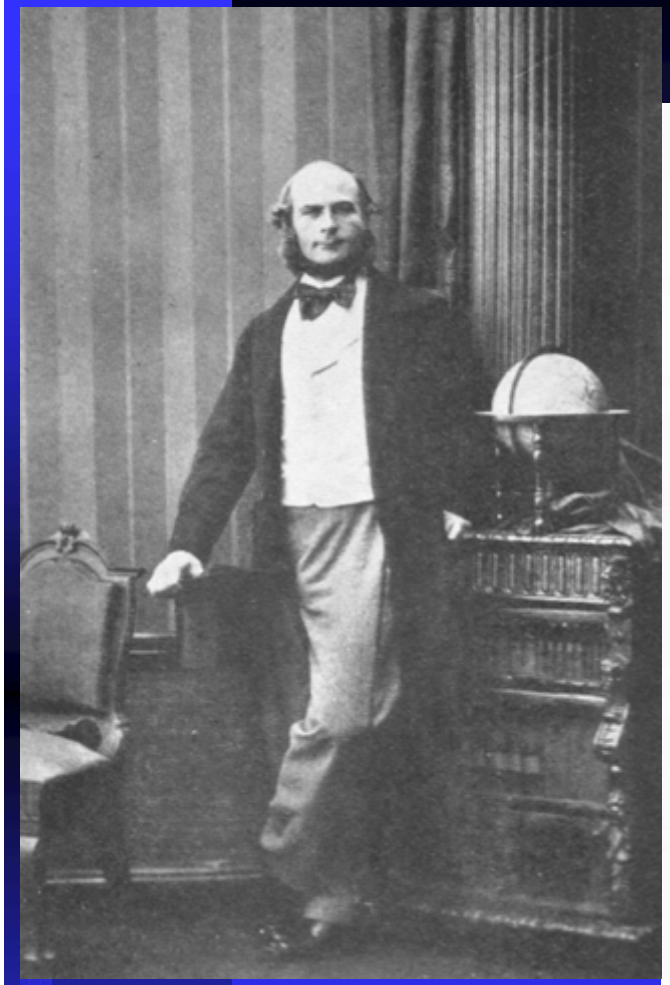
We would like to quantify day to day variations in predictability…

2 Feb 2006

Predictability

In our models or of the real world?

(Models first, just to get our terms defined; but we so not want to stay in the perfect model scenario!)



This Galton Board is a mathematical model.

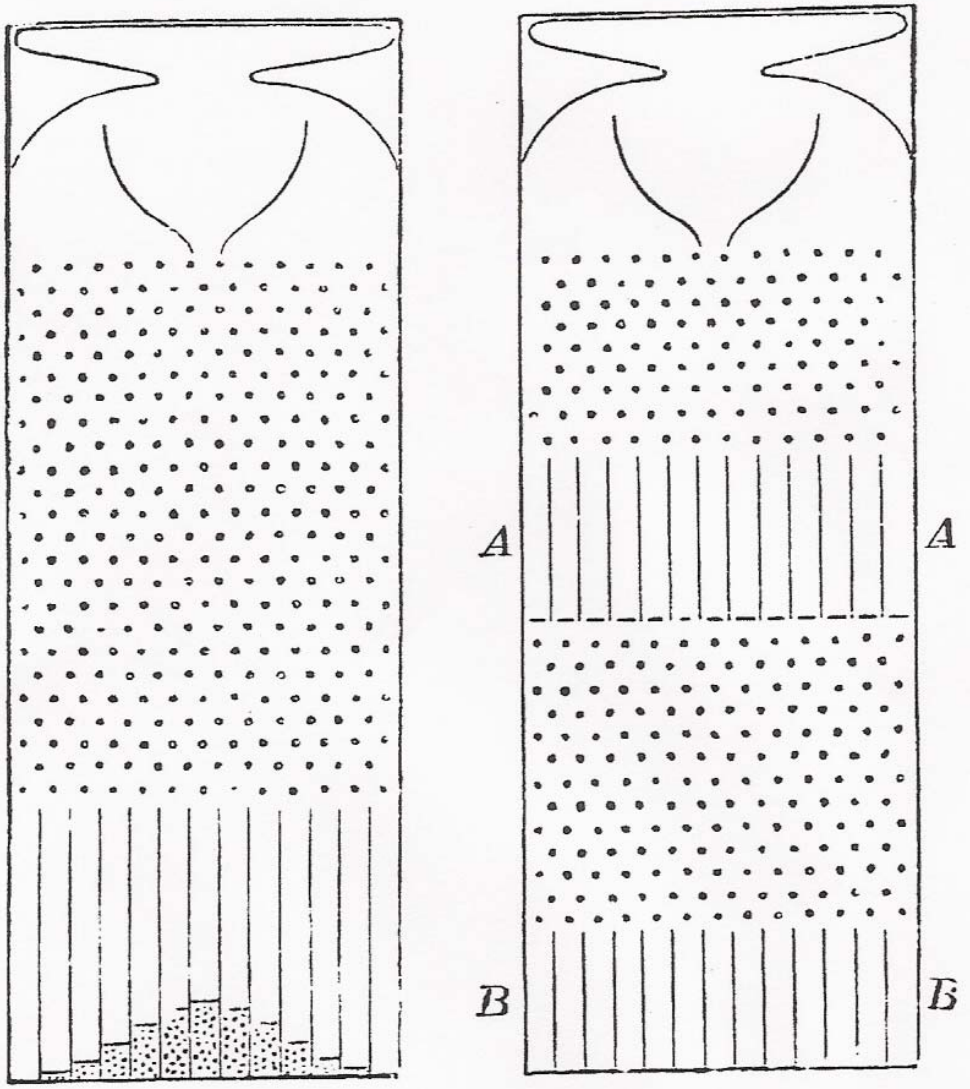
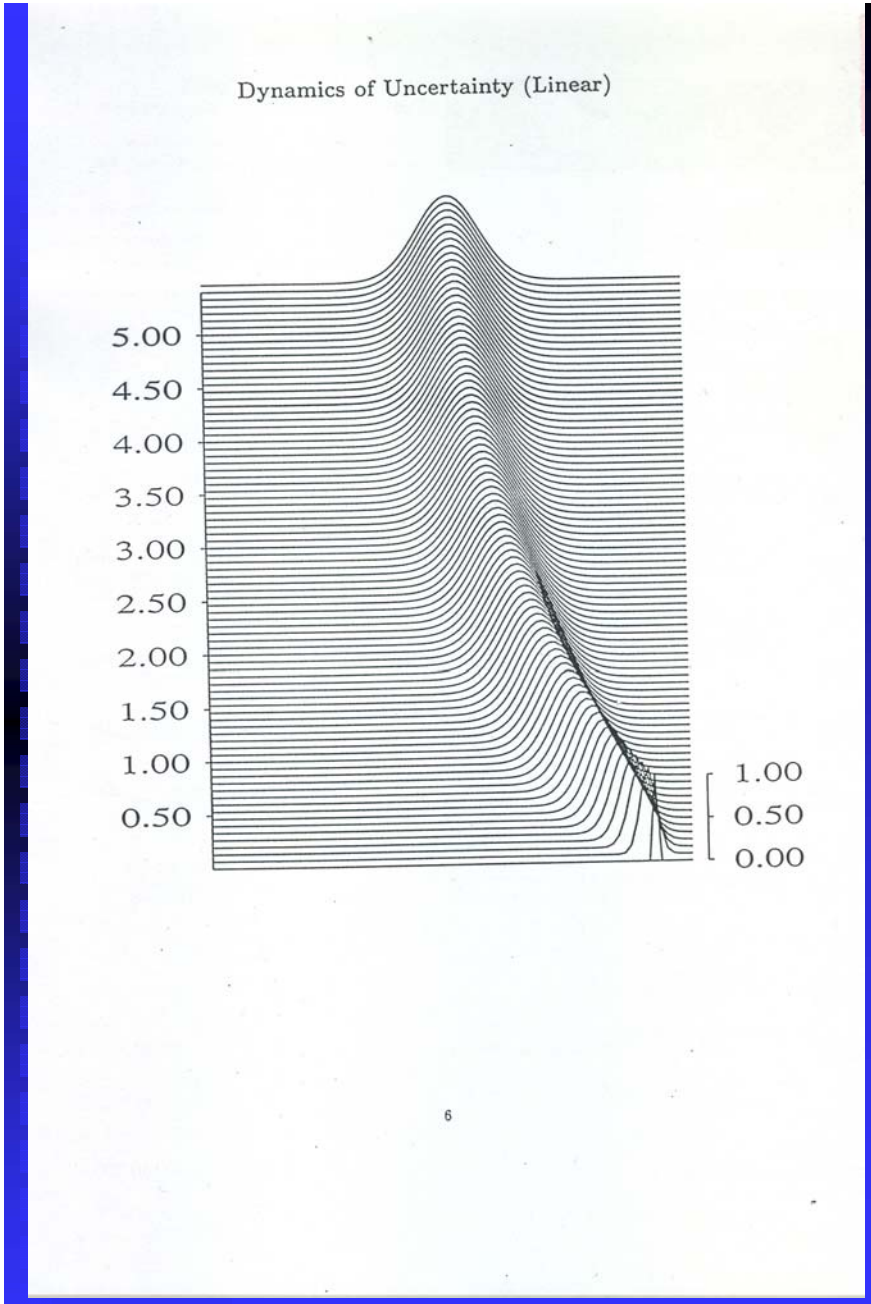
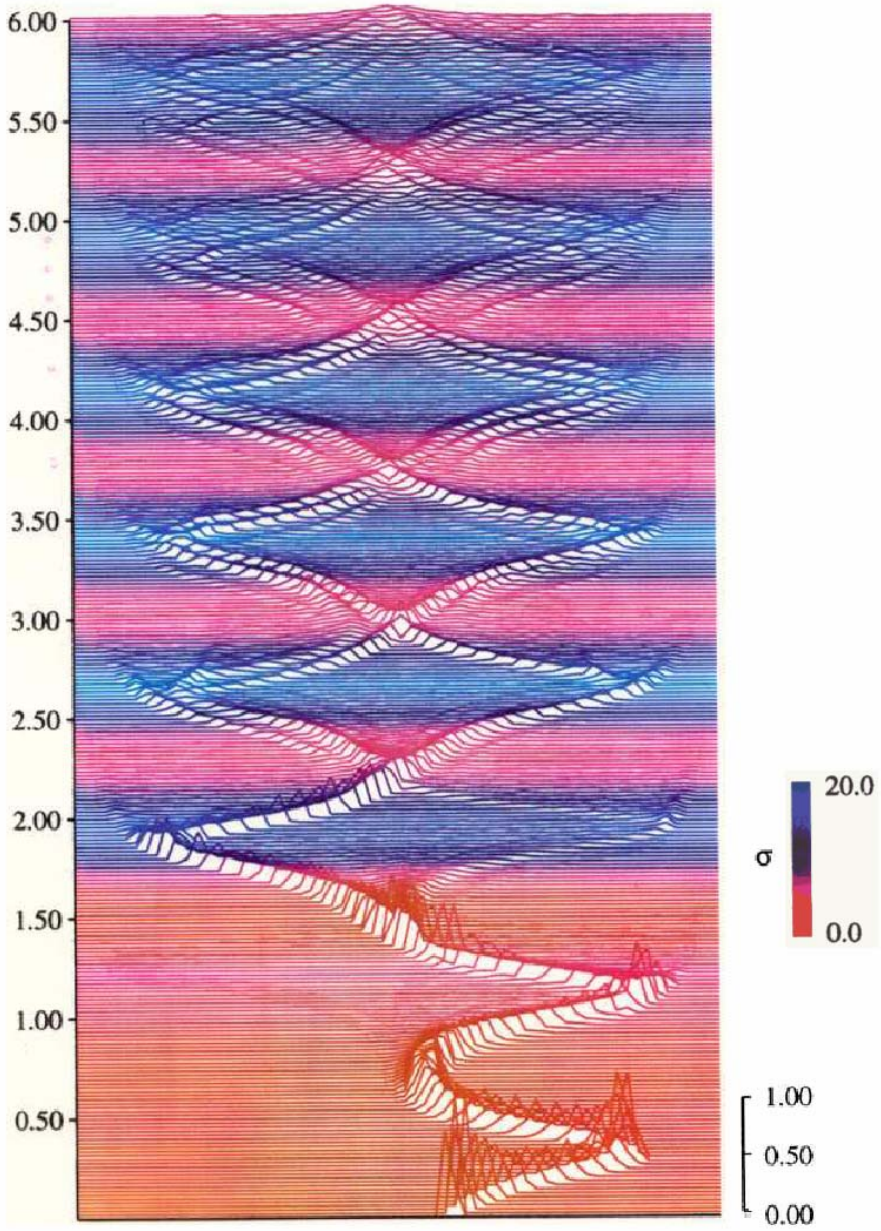


Figure 9.2 A schematic drawing of Galton's Quincunx, from Galton (1889a, p. 63).

2 Feb 2006



2 Feb 2006



Smith (2002) Chaos and Predictability in *Encyc Atmos Sci*

Ideally, we'd like a probability forecast…

…since observational limitations imply that, even with a perfect deterministic model, *the future* is at best a probability density function.

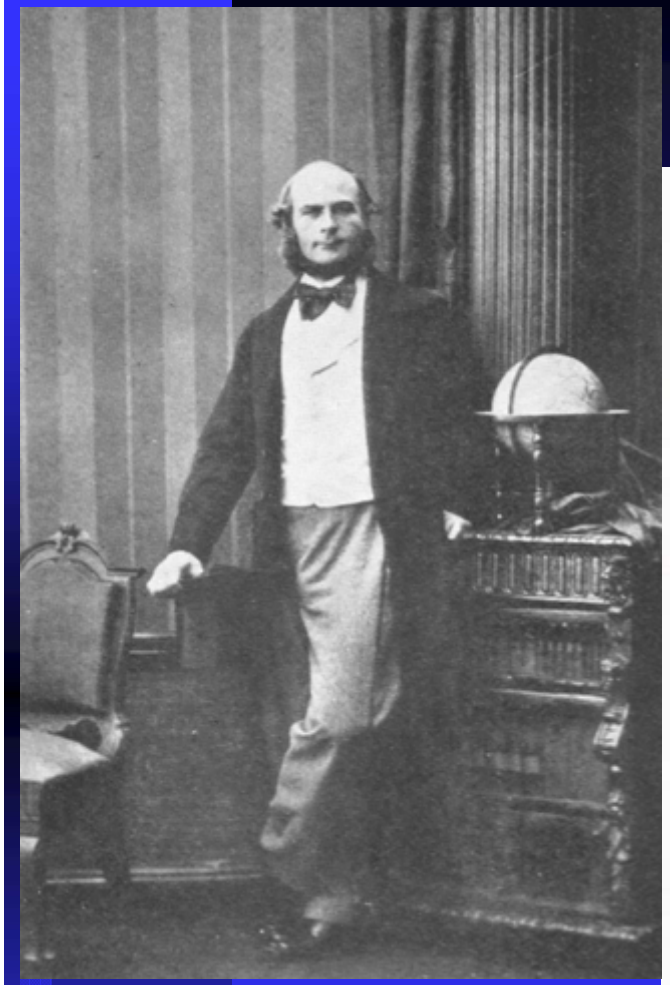
How do we translate finite monte carlo samples into PDF? (51 points in 10^7 d)

Note that RMS forecast error is at best irrelevant. (McSharry & Smith, PRL, 1999)

What skill scores should we be using?
Proper? Local?
(Ignorance? Good, 1952; Roulston & S 2003)

How can we best combine existing multi-model output (given that each model is inadequate)?

How do we pick the 51 points in 10^7 d?



2 Feb 2006

I term this a thought experiment because, while Galton clearly in several places described the variant of the Quincunx that performed the experiment, there is no indication that he actually built the apparatus. And having tried to build such a machine, I can testify that it is exceedingly difficult to make one that will accomplish the task in a satisfactory manner.

Stigler, 1999

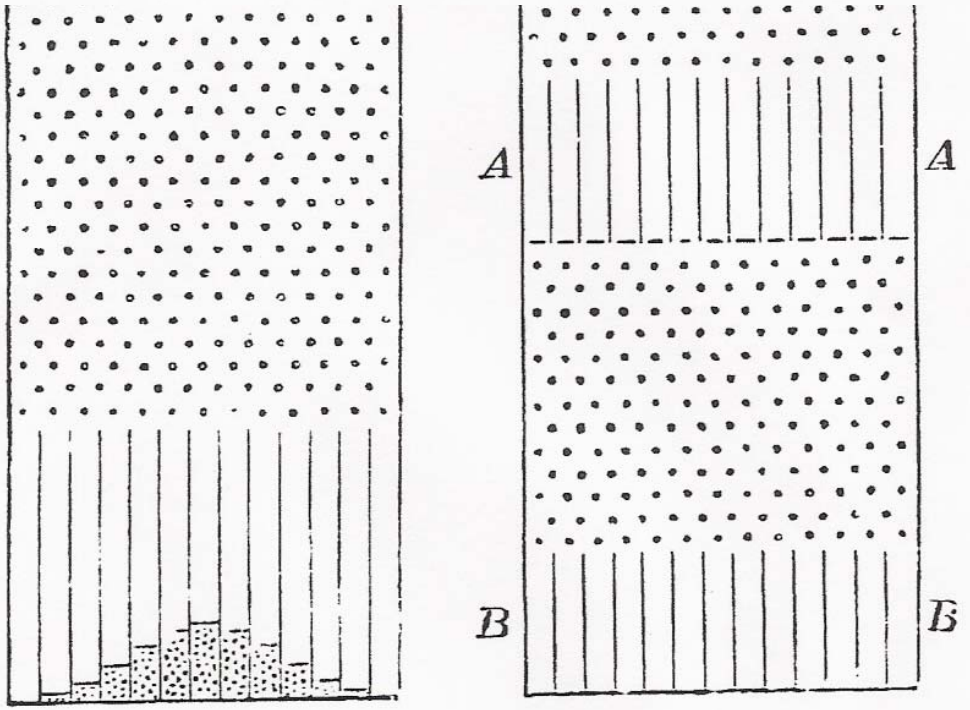
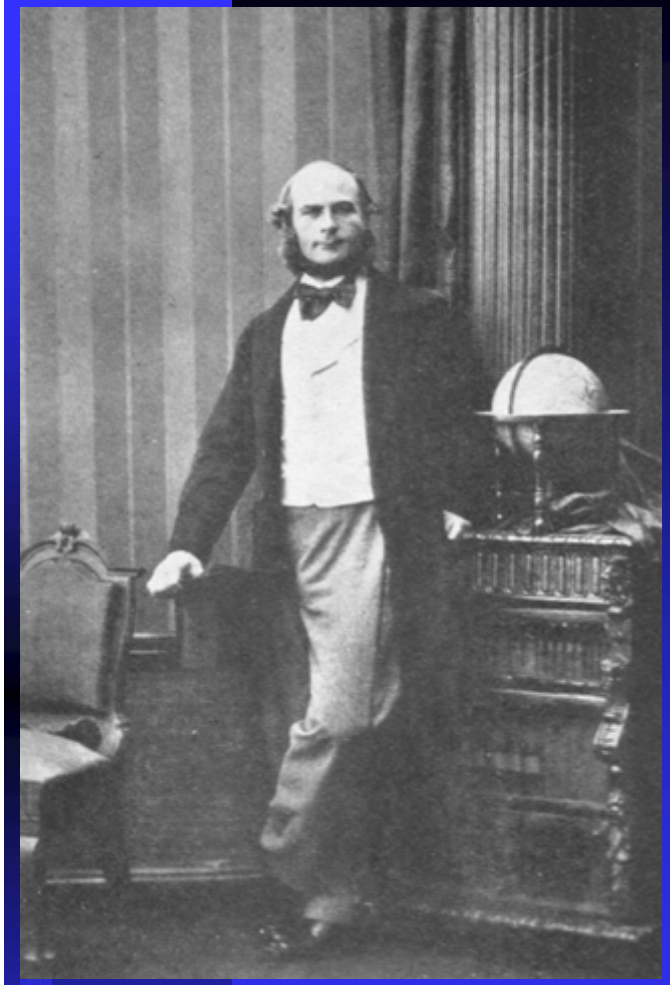


Figure 9.2 A schematic drawing of Galton's Quincunx, from Galton (1889a, p. 63).



I term this a thought experiment because, while Galton clearly in several places described the variant of the Quincunx that performed the experiment, there is no indication that he actually built the apparatus. And having tried to build such a machine, I can testify that it is exceedingly difficult to make one that will accomplish the task in a satisfactory manner.

Stigler, 1999

While this is Not A Galton (NAG) Board.
It is neither stochastic or chaotic; but at least it is!

2 Feb 2006

This is a NAG Board

The NAG Board (Not a Galton Board)
2 Feb 2006

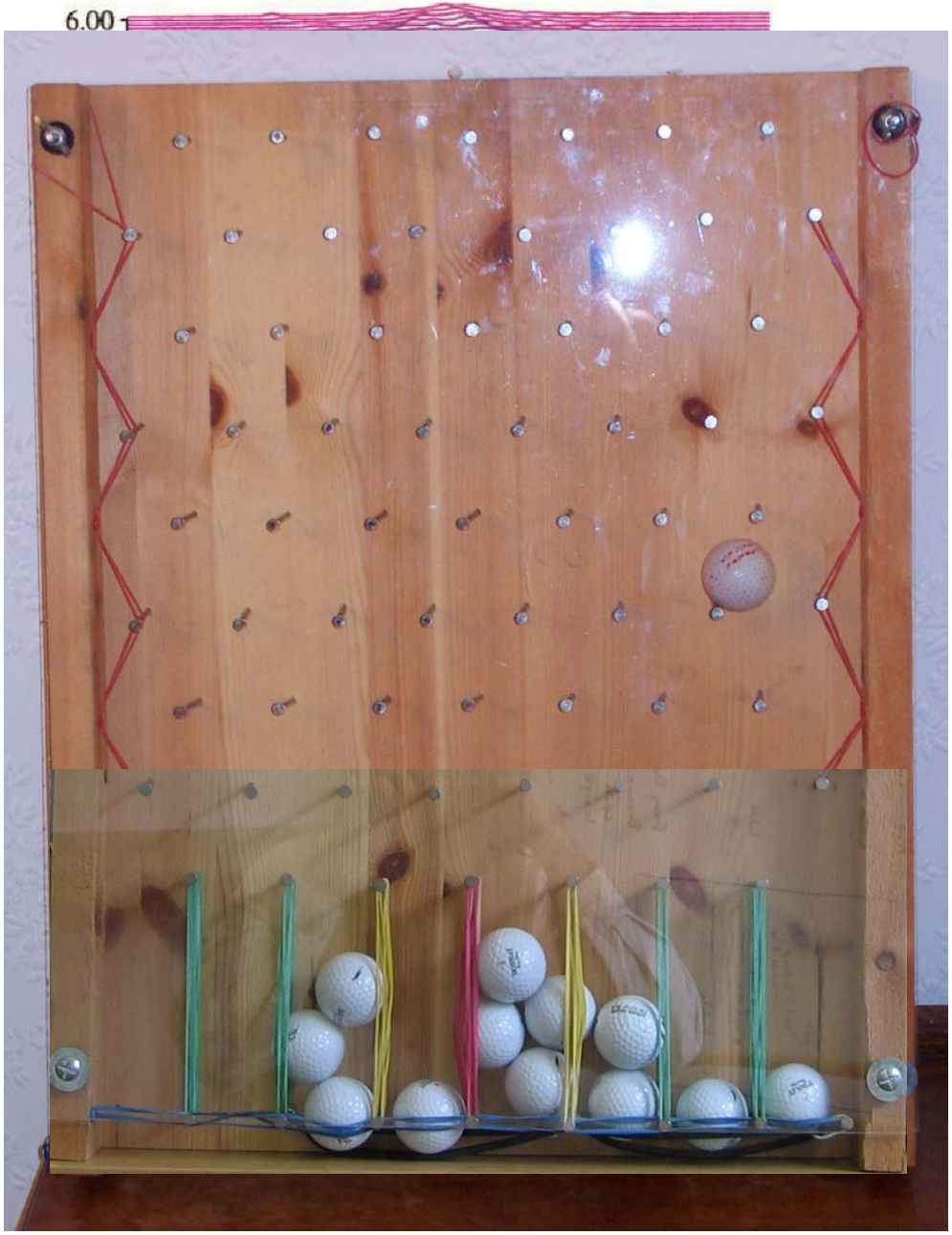
Uncertainty in the NAG board corresponds to predicting with a collection (ensemble) of golf balls…

Ensembles inform us of uncertainty growth *within our model!*

But reality is not a golf-ball; this EPS must deal with model inadequacy.

Nevertheless, weather EPS are useful!
Operational Day ~10 Weather Ensembles:
US and European Services: 1992
Canada: Now

Lorenz PDF evolution



from: Encyclopedia of Atmos Sci (2003) Wiley

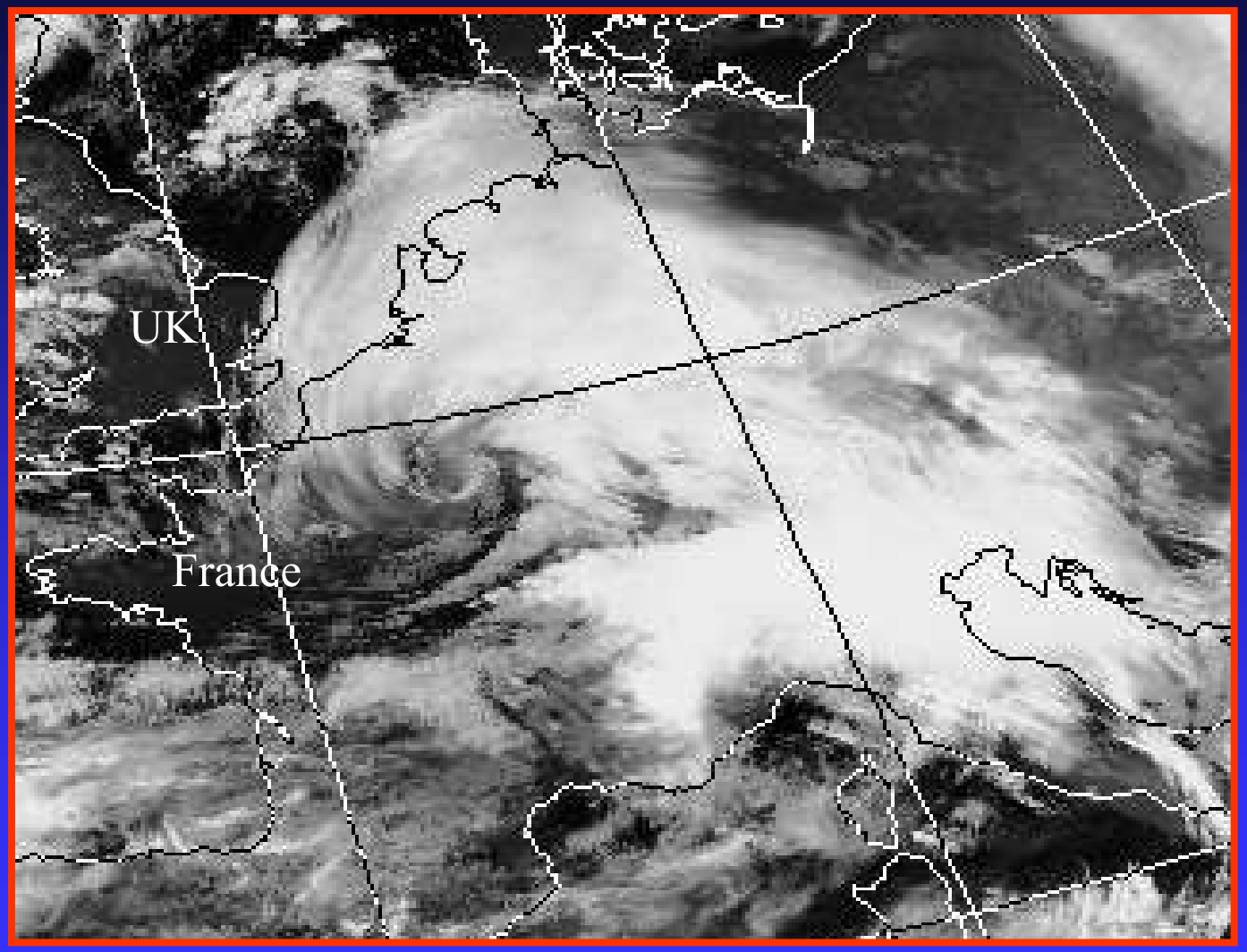
The ensemble message has been taken to heart within meteorology, where ensemble forecasts have been operational for over a decade.

(In both Europe and America)

In the NAG board, this corresponds to predicting with a collection (ensemble) of golf balls…

What does this have to do with NWP?!?

Could we have better seen Lothar coming?



Dundee Satellite Station: 0754 UTC 26 December 1999

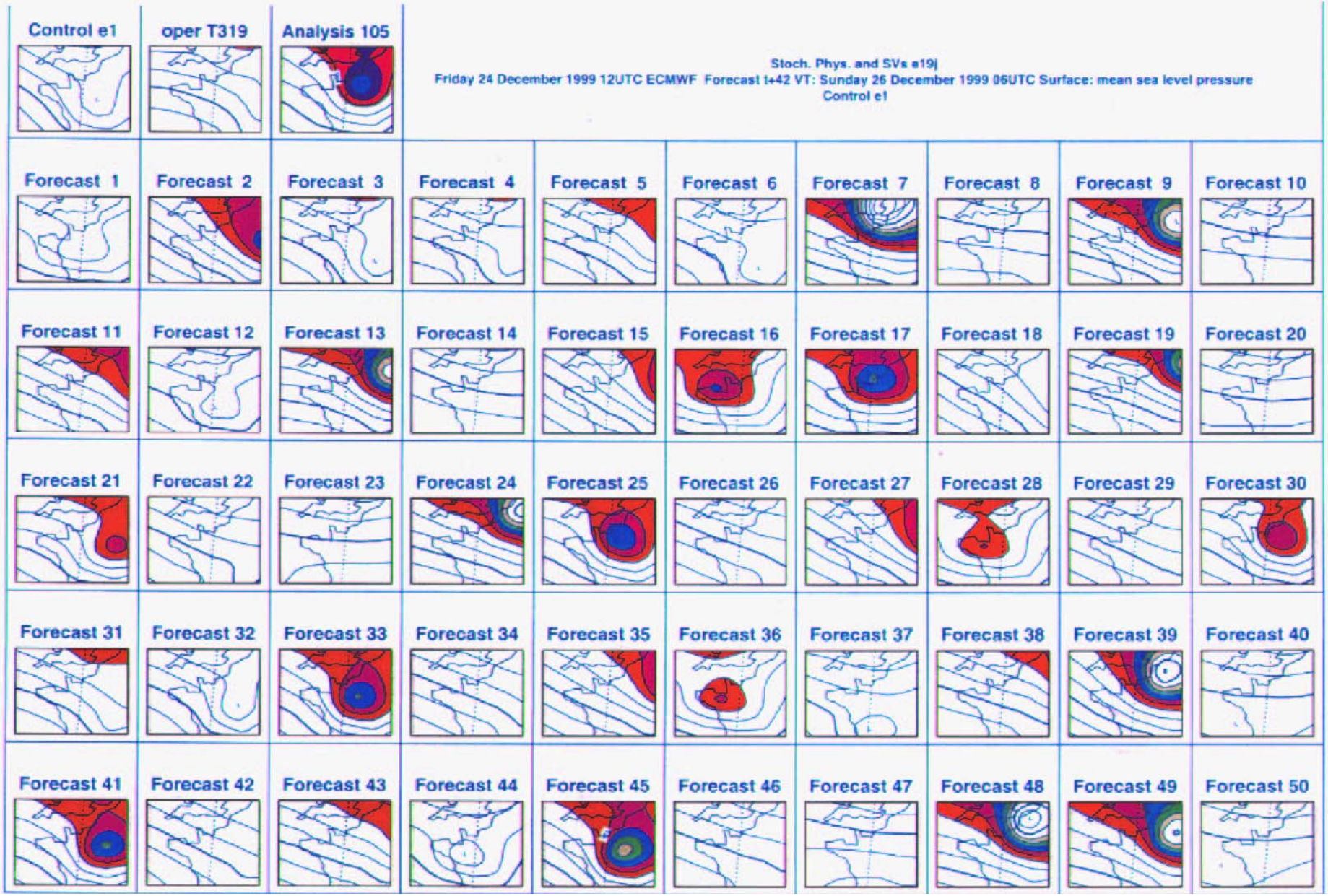
2 Feb 2006

Two 2-day forecasts and the weather for December 26 1999

24 December 1999 12UTC ECMWF Forecast
6 December 1999 06UTC Surface: mean sea l
Stoch. Phys. and SVs e19j

The single forecast from the "best" model

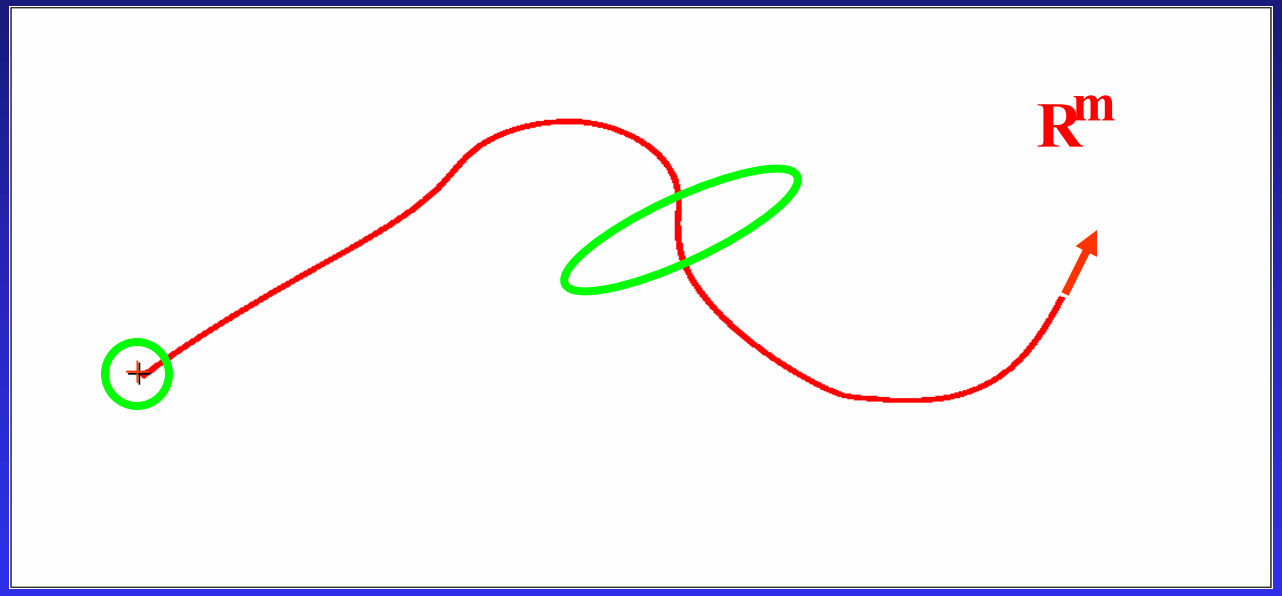
These model runs were launched simultaneously in the leading 25D SV space. These are a collection of golf balls! (not a relevant probability!)



Where did these forecasts come from?

How do we pick the 51 points in 10^7 d?

Suppose we have already estimated the current state of the atmosphere



…and a trajectory (forecast) from that model initial condition. Under a smooth dynamic, an infinitesimal sphere of uncertainty will evolve into an ellipse; we might sample preimages of the "leading" axes of this ellipse.

In the mid 1960's, Lorenz suggested this approach, by sampling the leading SV subspace:

The evolution of an infinitesimal uncertainty over a finite time Δt is determined by the linear propagator $M(\mathbf{x}_0, \Delta t)$ along the trajectory $\mathbf{x}(t)$, that is

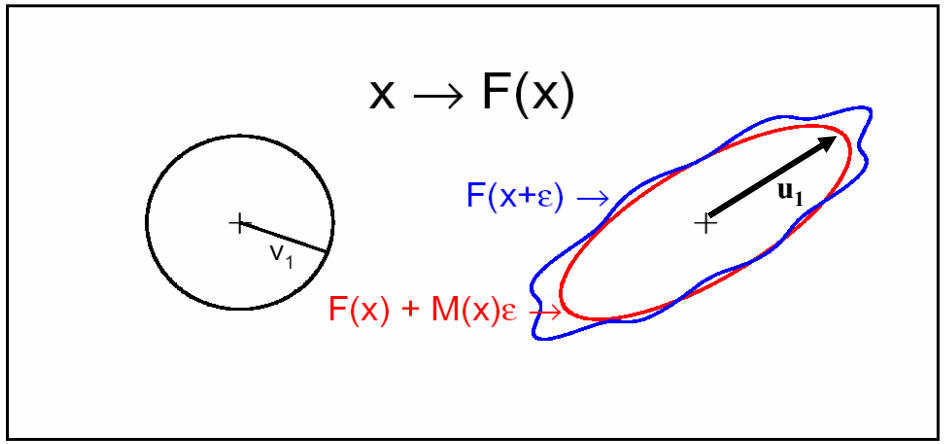
$$\boldsymbol{\epsilon}(t_0 + \Delta t) = M(\mathbf{x}_0, \Delta t)\boldsymbol{\epsilon}(t_0) \tag{7}$$

where $\mathbf{x_0} \equiv \mathbf{x}(t_0)$ and, for a flow,

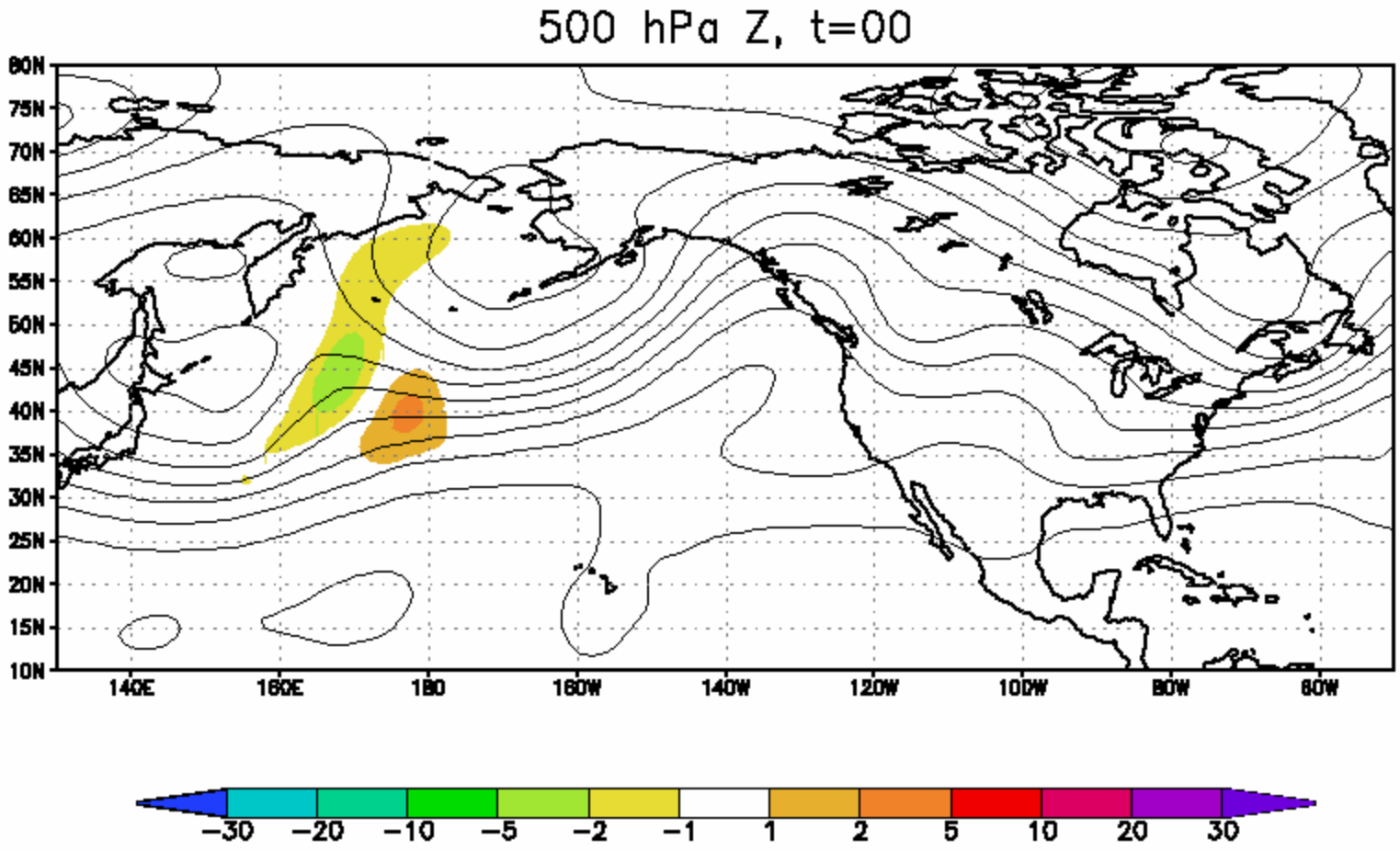
$$M(\mathbf{x_0}, \Delta t) = \exp\left(\int_{t_0}^{t_0+\Delta t} J(\mathbf{x}(t))dt\right). \tag{8}$$

For discrete time maps, the linear propagator is simply the product of the Jacobians along the trajectory

$$M(\mathbf{x_0}, k) = J(\mathbf{x}_{k-1})J(\mathbf{x}_{k-2})\ldots J(\mathbf{x}_1)J(\mathbf{x}_0). \tag{9}$$

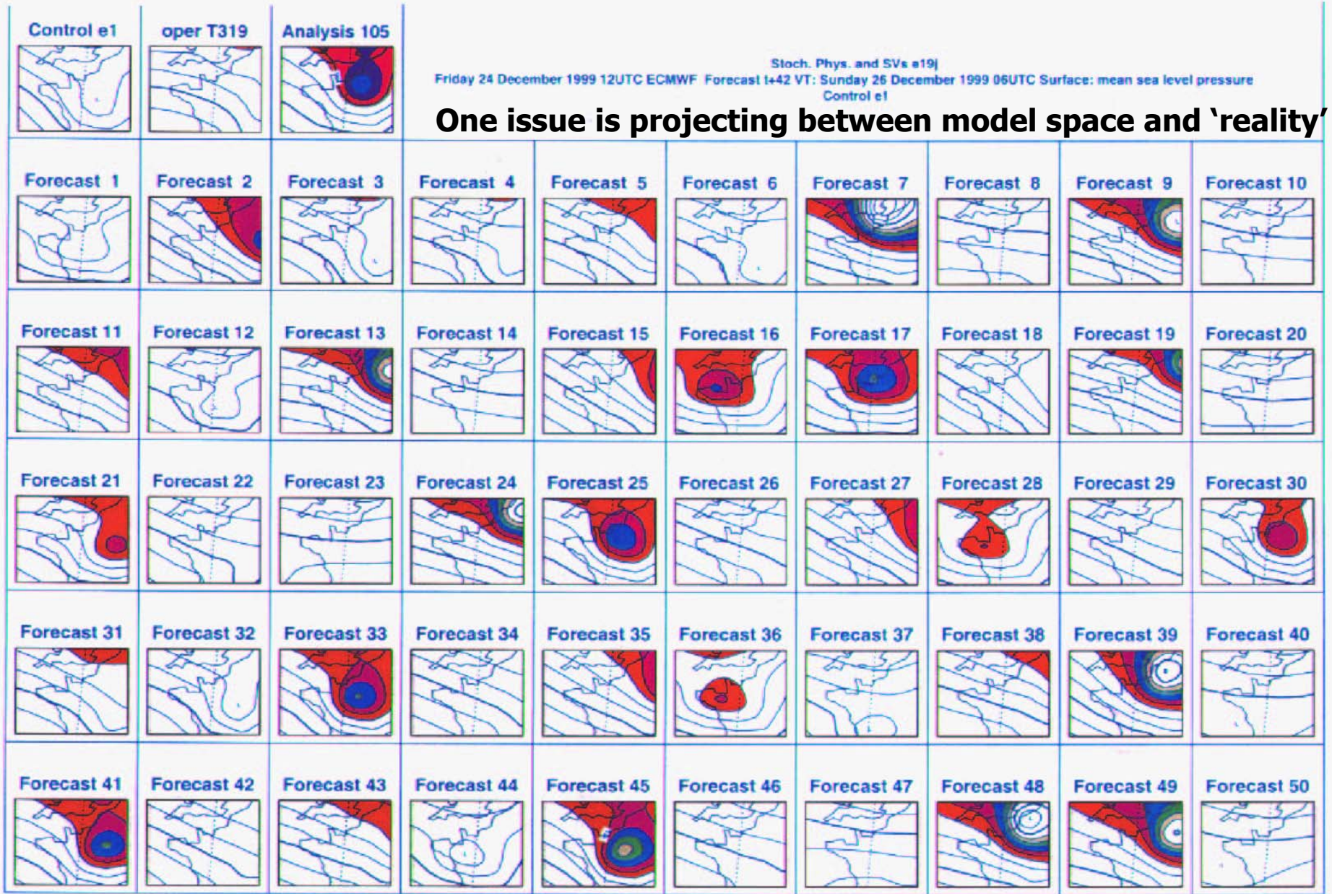


This is (a slice of) v1 of an operational NWP model. And this is (a slice of) u1 (3 days later)



NOGAPS slides from Carolyn Reynolds

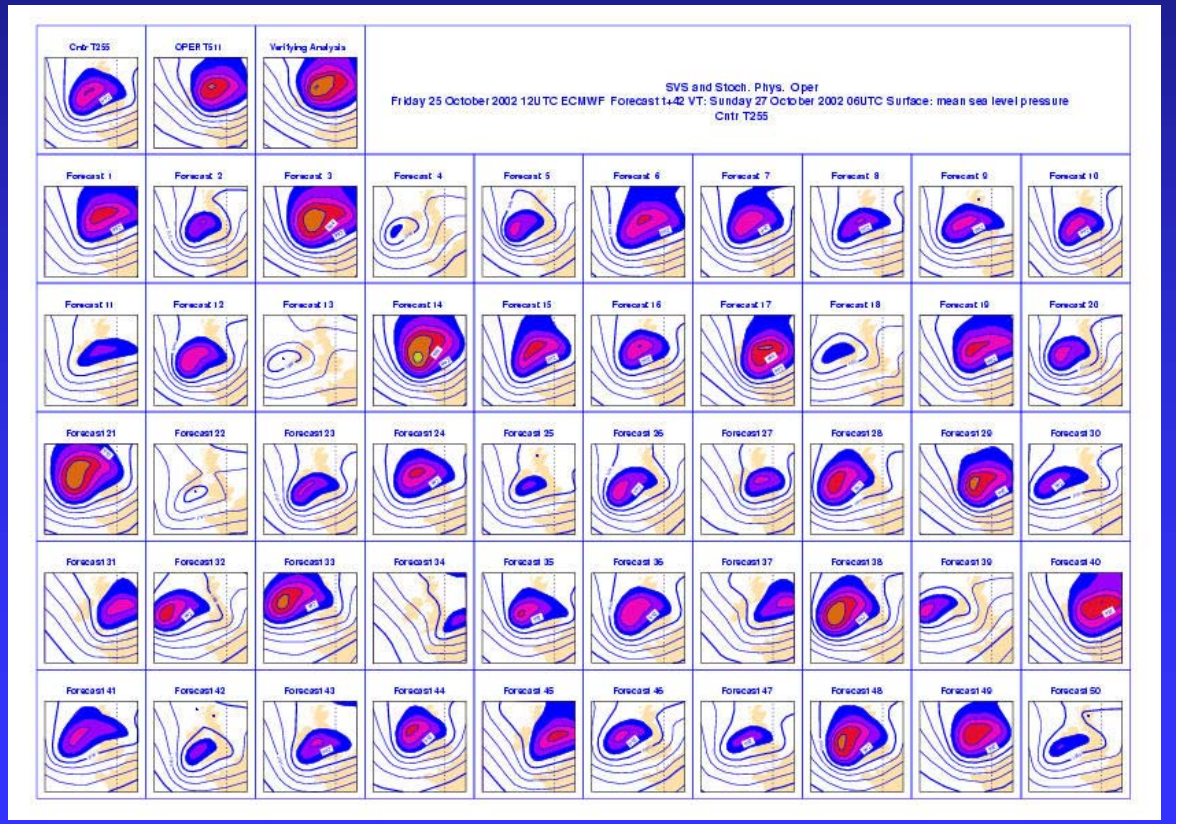
These model runs were launched simultaneously in the leading 25D S space. How do we interpret these scenarios?



One issue is projecting between model space and 'reality'

Of course, early warning would not have saved the trees of Versaile, forecasts are of "value" only if mitigation is an option, given early warning.

This is the 42 hour ECMWF ensemble forecast for the Oct 22 2002 storm over England; even if I do not know how to turn this samle into a probability forecast, I do know that I do not want my car parked under a tree…



2 Feb 2006

The value of a forecast depends on our ability to act under uncertainty.

2 Feb 2006

SVs depend on the metric(s) used to define them, suggesting they can be used to target dynamically important "sensitive regions" for observation.

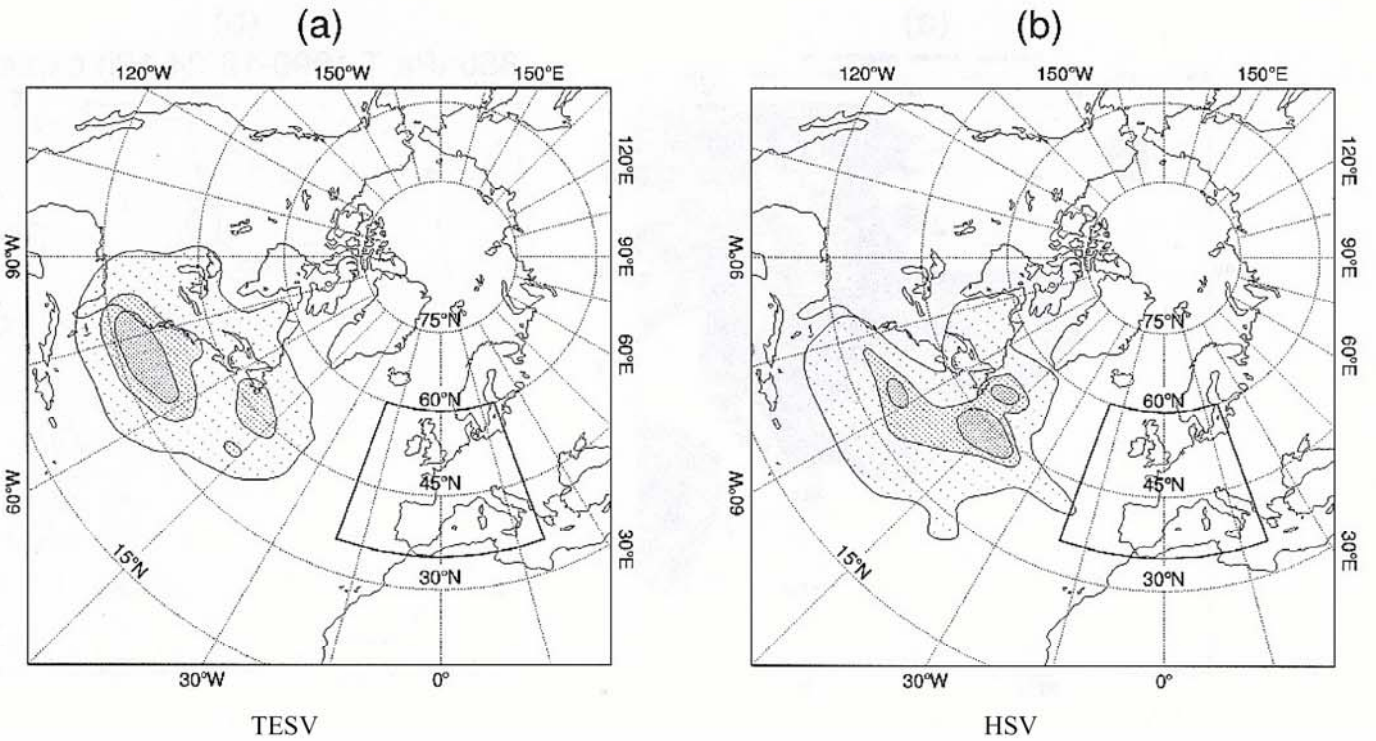
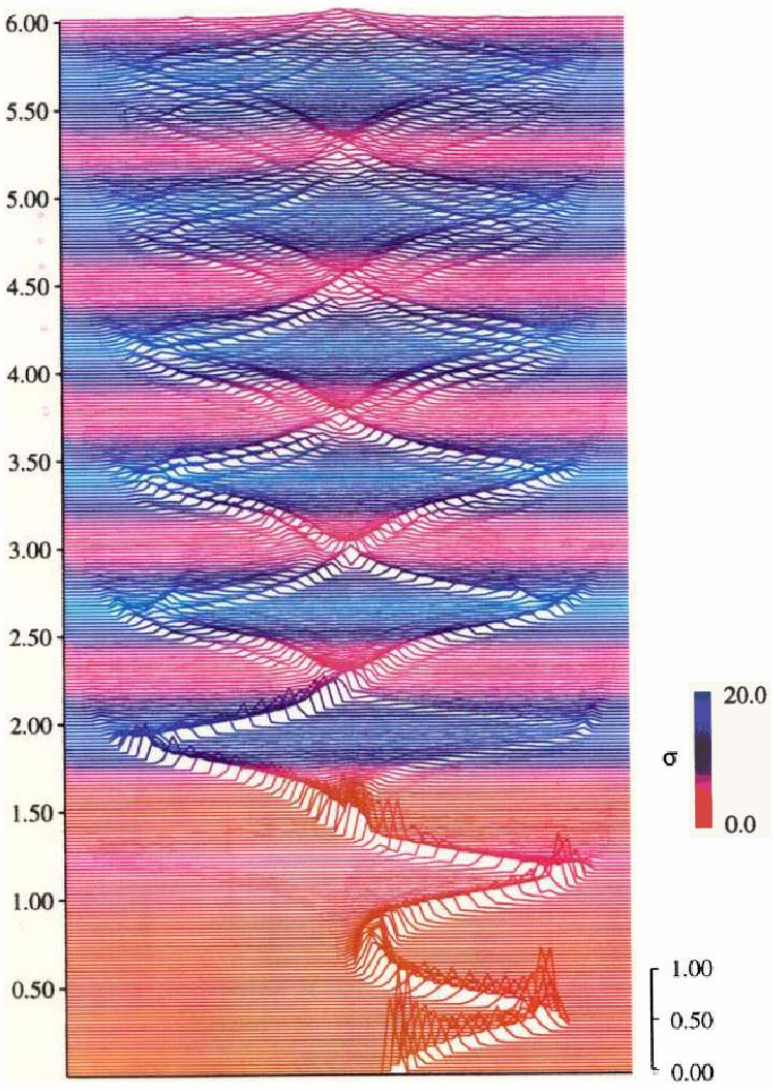


Figure 1: Sensitive regions for Lothar predicted based on (a) total energy and (b) Hessian singular vectors. The box indicates the verification area. From Leutbecher et al. (2002).

Good methods for selecting which obs to *ignore* will be needed in a year or two.



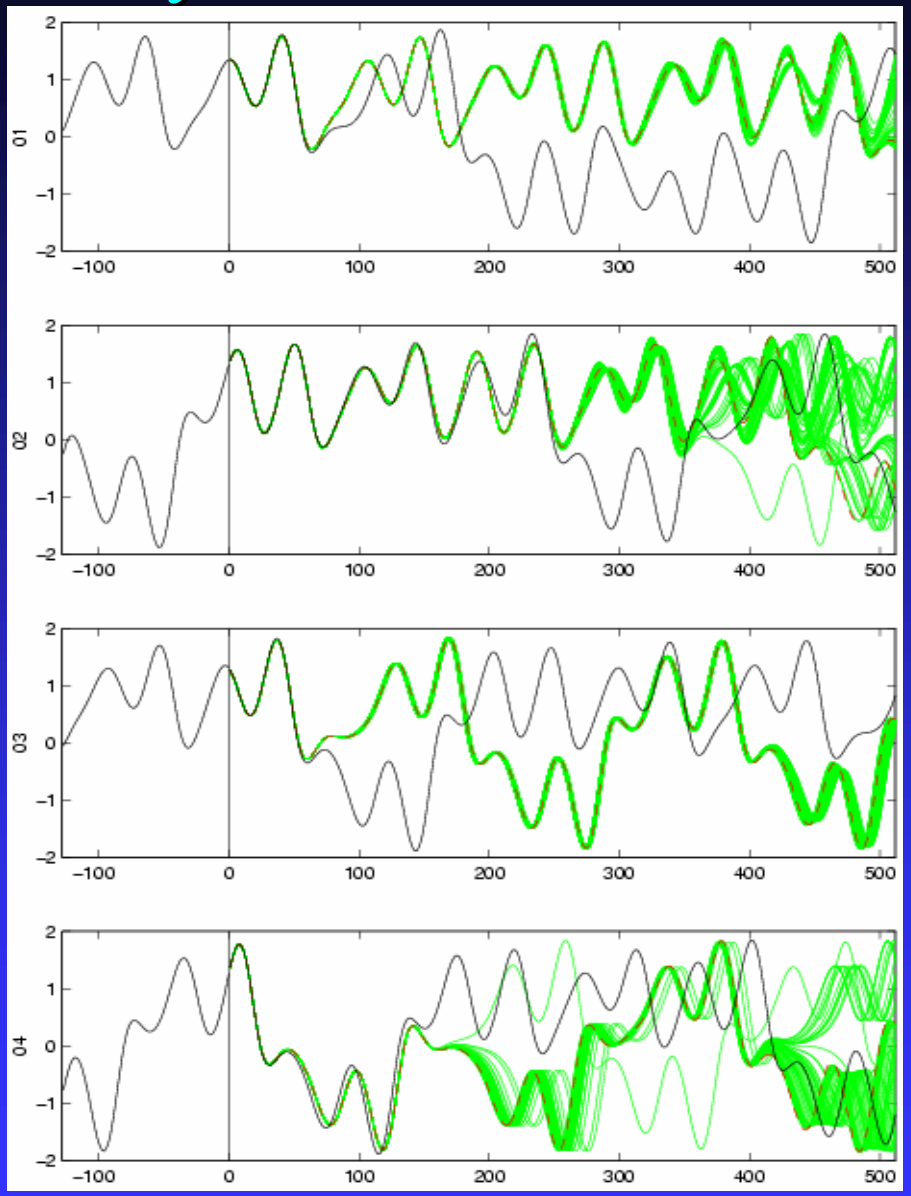
But how does one convince a Laplacian meteorologist that, given a finite amount of cpu, it makes more since to run a lower resolution model many times, than a high resolution model once?

And how might you determine the trade-off between ensemble size and model resolution?

What skill scores should I use for probability forecasts generated from *imperfect* models?

In short: how can we evaluate an EPS?

Why is there a trade-off: Chaotic Circuit



Short term (weather) forecasts are very skilful but model inadequacy leads to poor model-based probability forecasts (and eventually from model irrelevance!)

A model can add value as long as it adds information, it need not have traditional "skill."

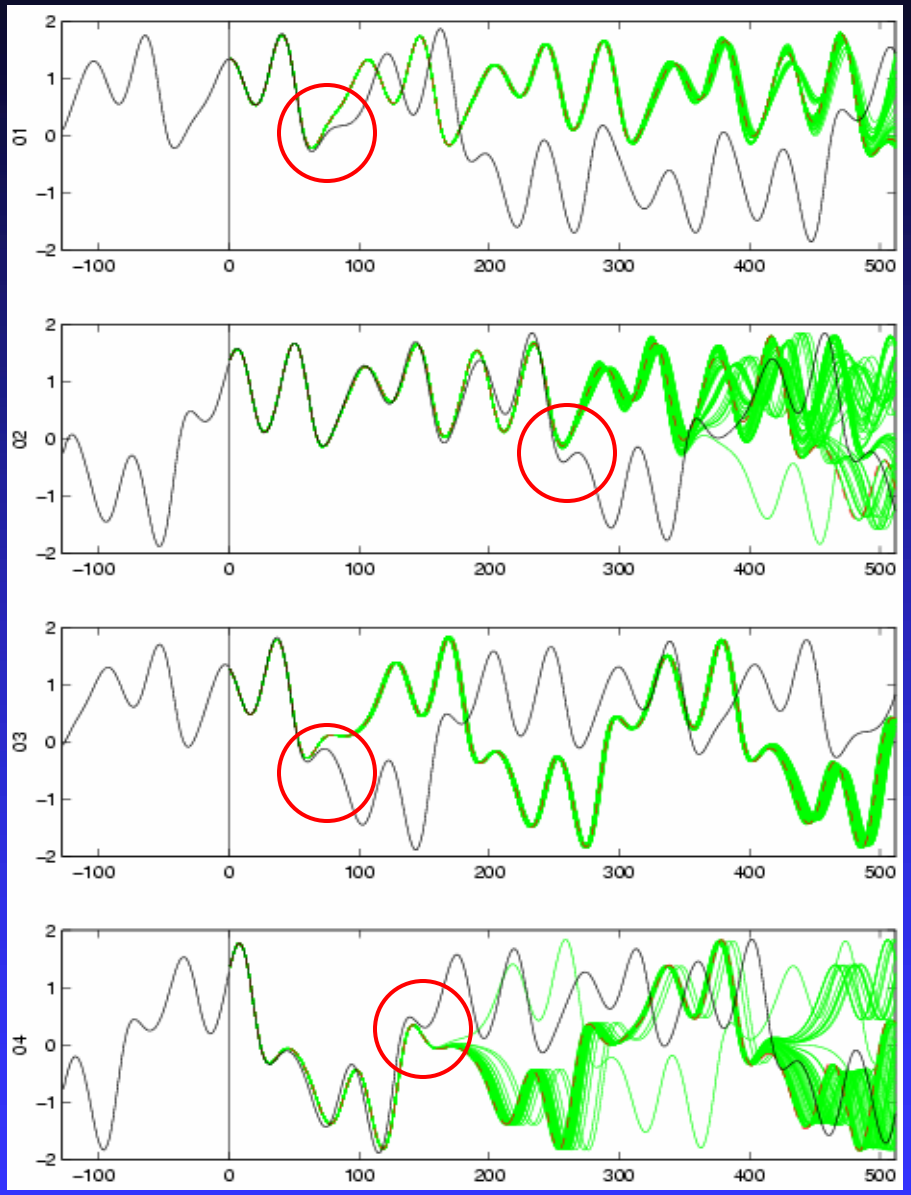
But how can we know how long an ensemble is likely to have at least one member that "resembles" the verification?

ensemble size vs complexity

512 member ensembles
Best known 1-step model
512 step free running forecasts

2 Feb 2006

Forecasts busts in a Chaotic Circuit



512 member ensembles
Best known 1-step model
512 step free running forecasts

We might wait until we know the future, then look for model trajectories that "shadow" the obs to within the noise.

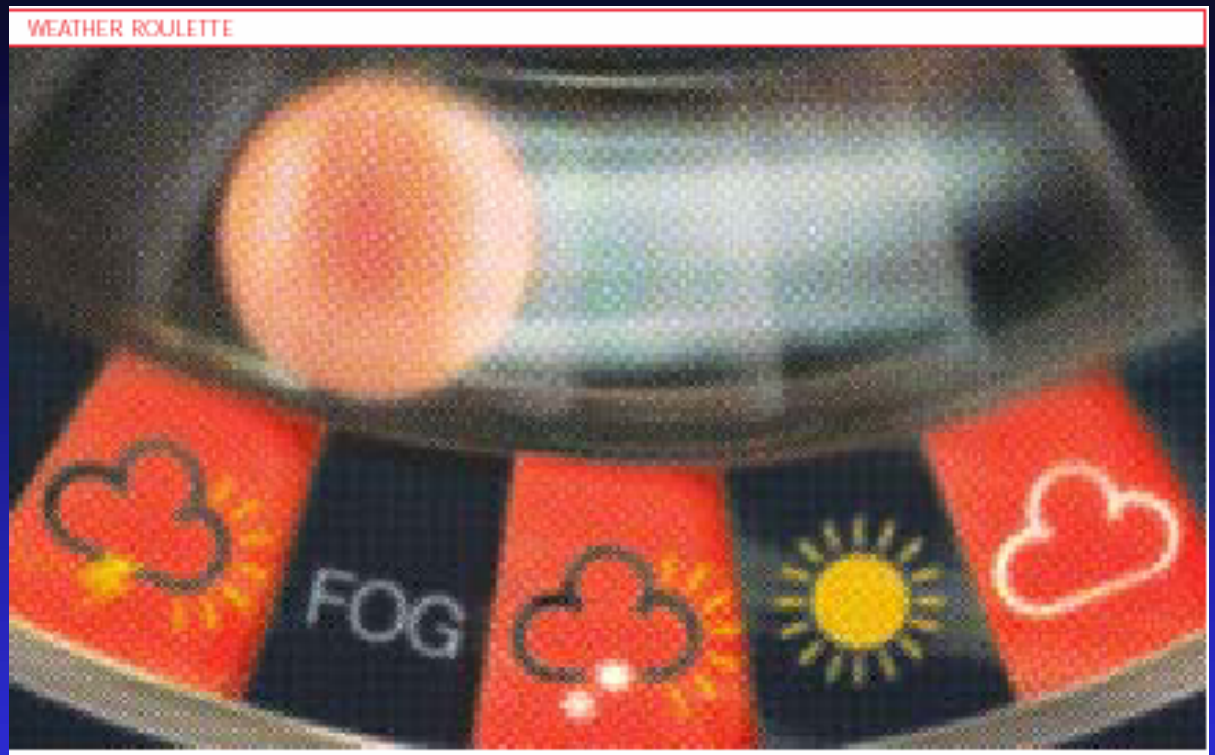
The distribution of shadowing times quantifies model inadequacy.

(But what is noise, really?)

State-dependent Systematic Model Error (in a recurrent system/model pair)

2 Feb 2006

In practice: Probability forecasts do not have to be accountable to be useful!



Weather roulette

Wager £100 each day on the temperature at Heathrow, betting an amount proportional to your predicted probability of that outcome (Kelly Betting).

How would a probability forecast based on the ECMWF EPS fare against a house that set its odds using climatology?

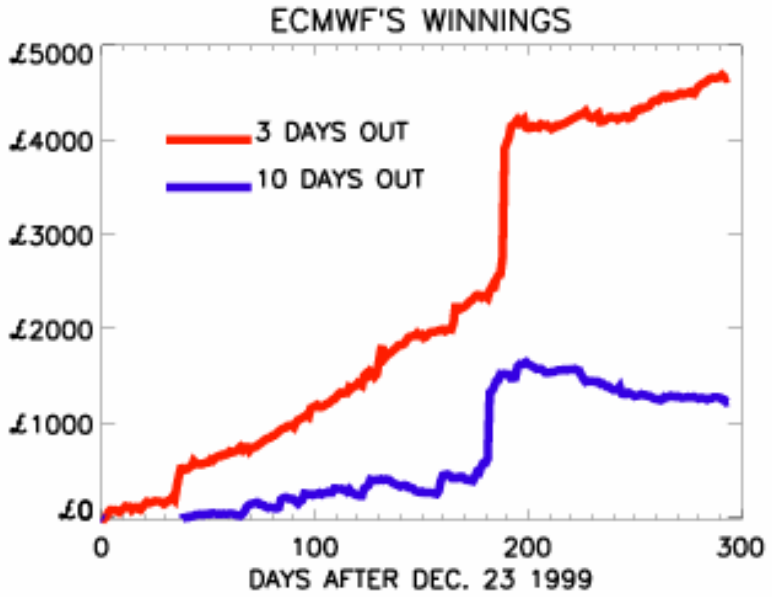
2 Feb 2006

WEATHER ROULETTE

TEMPERATURE AT HEATHROW

TABLE MAXIMUM: £100

1982–99 CLIMATOLOGICAL ODDS



TEMPERATURE (°C)

25	26	27	28	29
20	21	22	23	24
15	16	17	18	19
10	11	12	13	14
5	6	7	8	9
0	1	2	3	4
−5	−4	−3	−2	−1

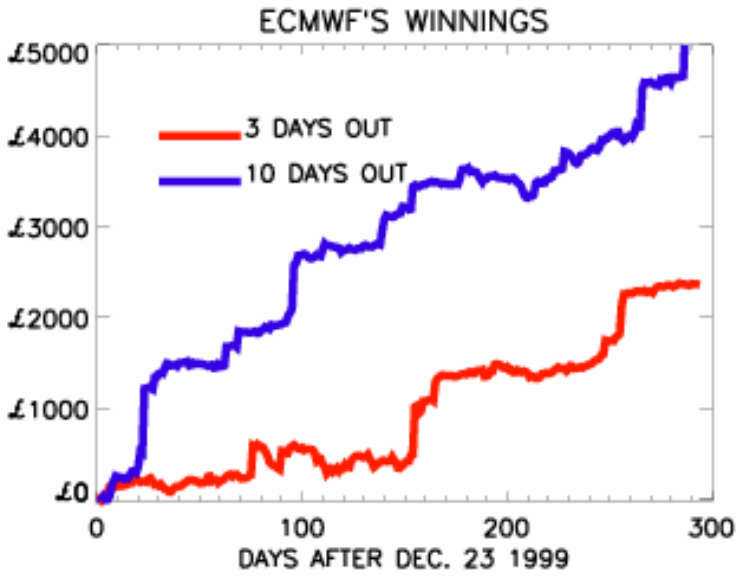
WEATHER ROULETTE

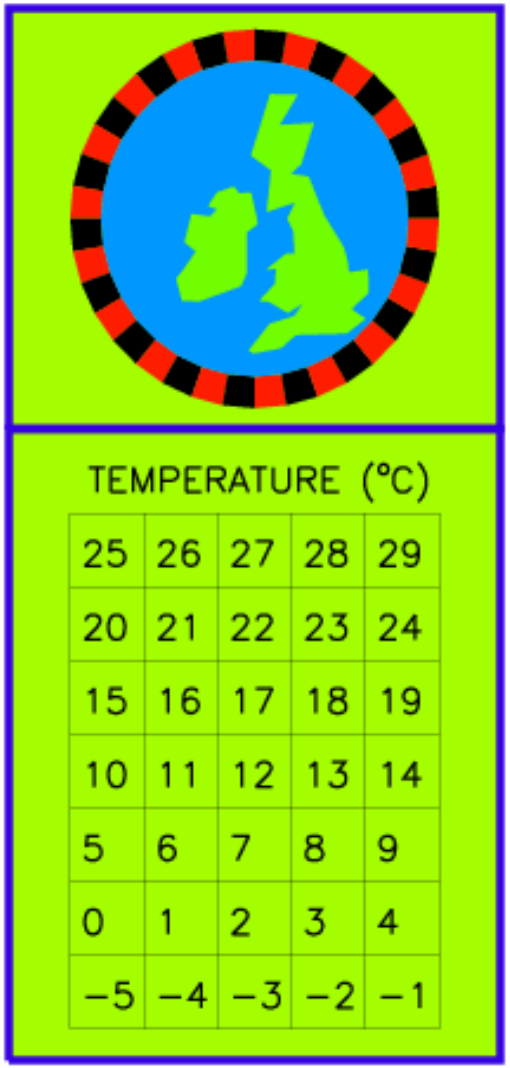
TEMPERATURE AT HEATHROW

TABLE MAXIMUM: £100

ODDS SET BY HIGH RES. FORECAST

BETS PLACED ACCORDING TO ENSEMBLE





Dressing allows a fair comparison of EPS and BFG.

How can we measure this kind of skill?

Dressing and Scoring

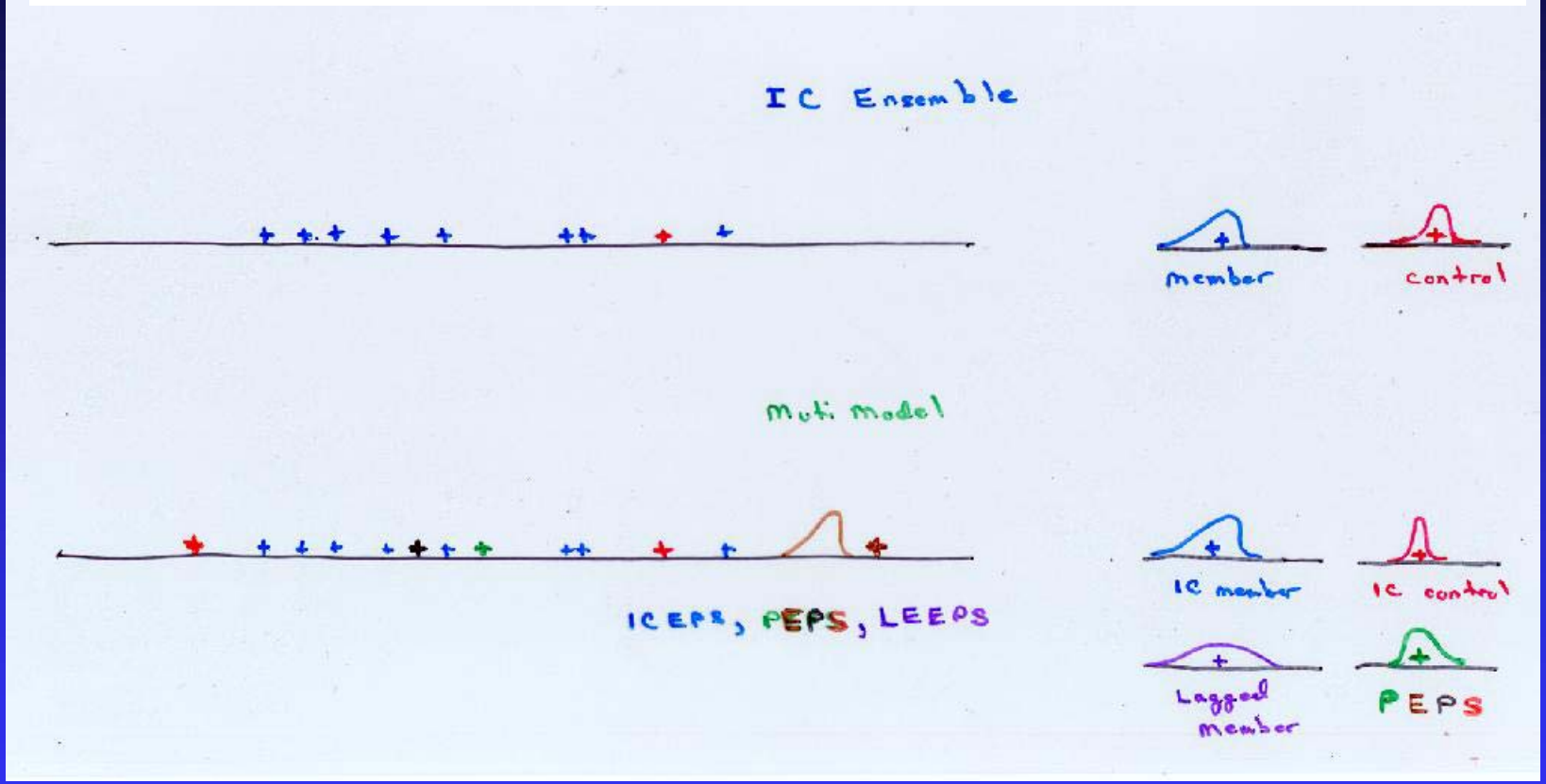
Ensemble forecasts in the model-state space can be *interpreted* as probability forecasts in the target space.

One way to do this is scenario dressing with kernels.

If done: how should these be evaluate? Tuned?

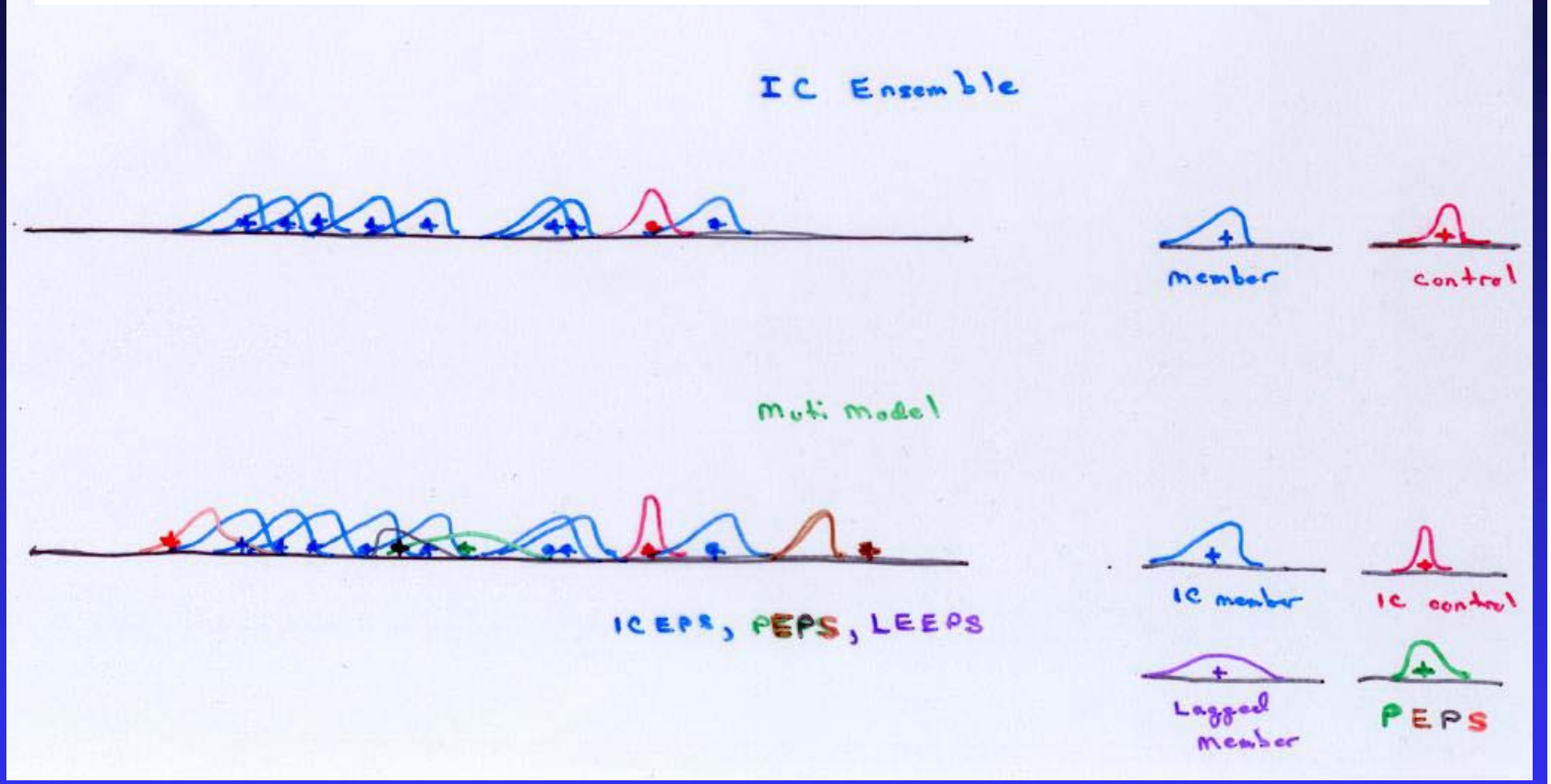
Is conditioning a probability forecast on the joint distribution of a multi-model ensemble feasible?

Dressing an Ensemble as a Collection of Scenarios:



Each class of ensemble member is dressed with its own kernel.
Members of an EPS, DEMETER, a PEPS or LEEPS are easily included.

Dressing an Ensemble as a Collection of Scenarios:



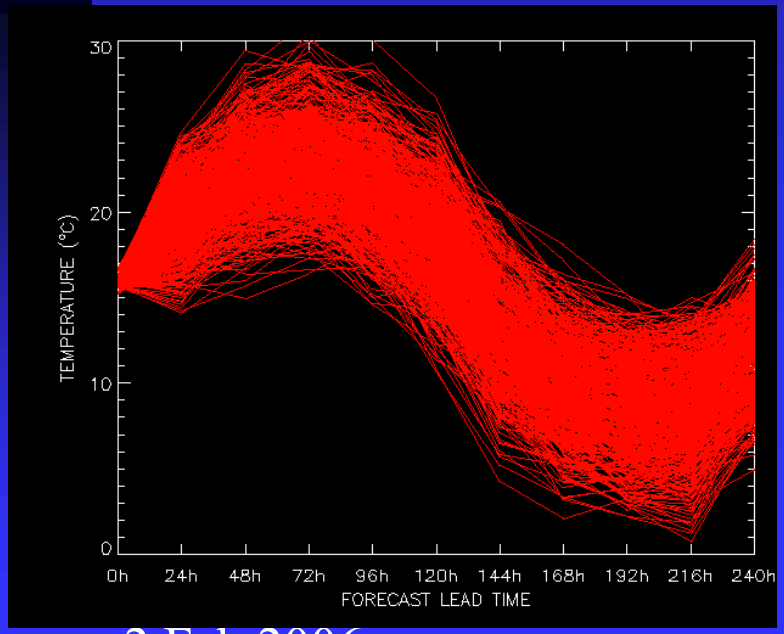
Although this yields a smooth distribution function, it is unlikely that a physical scientist would want to call it a probability forecast, as the models are (each) known to be imperfect *a priori*.

Bayesian model averaging fails to be internally consistent for the same reason.

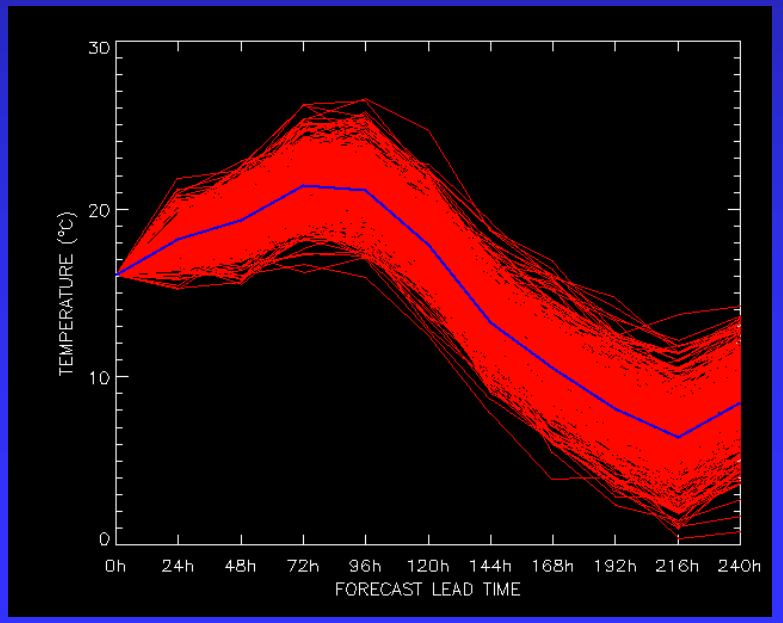
Reasons for Scenario Dressing:

- "Probability" Forecasts:
  - Accounting for finite ensemble size
  - Accounting for typical model inadequacy
- Equal Counting Statistics:
  - A fair comparison between EPS of different sizes/compositions
  - A fair comparison between an EPS and a single hi res BFG simulation

Would you rather play informed roulette with one $50 chip or 50 $1 chips?



100 Element Dressed EPS Forecast

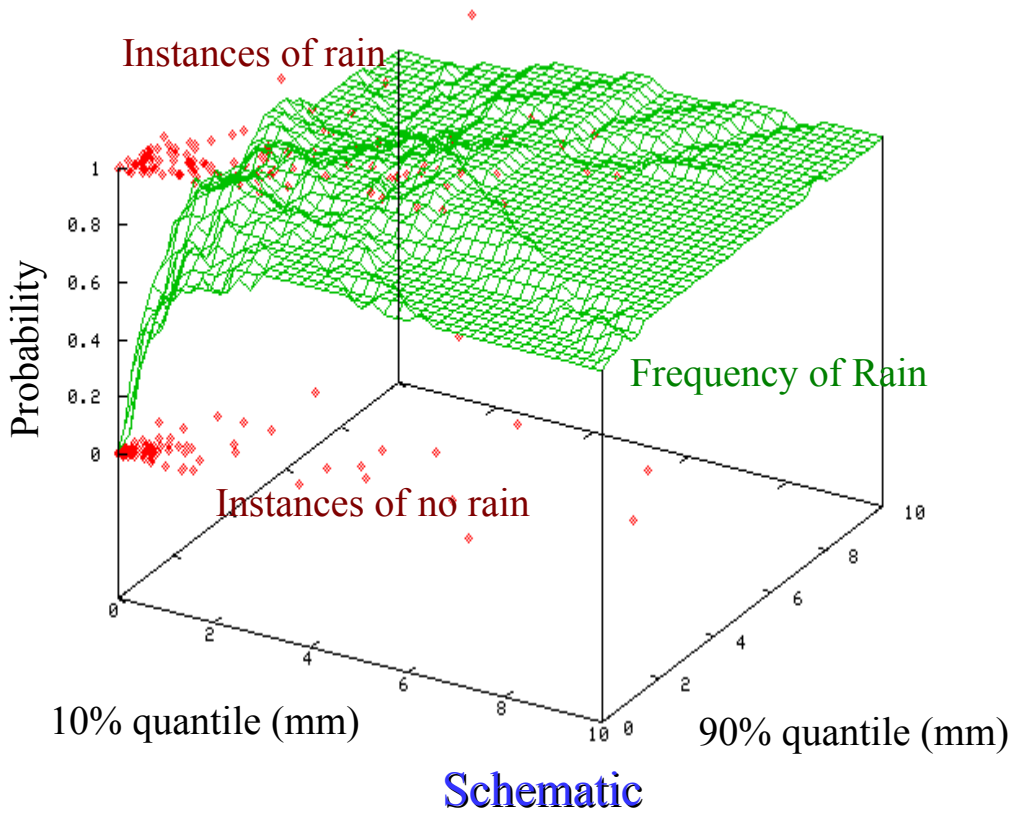


100 Element Dressed BFG Forecast

A final aside: Precipitation at Schleswig

We don't have to interpret model-rain *as* a scenario for rain in each ensemble member...

Better to use the joint distribution with the aim of extracting information.



Schematic

Eight Current Challenges:

Moving Beyond Scenarios
$P(\boldsymbol{x} \mid \text{obs}; X_1, X_2, X_3; Y_1, Y_2; Z_0)$

Dressing individual ensemble members may be useful, but a better (& more Bayesian) approach would be to condition on the joint distribution of our (imperfect) models.

Projection Operator –or- Ensemble "Bias Removal"

We do not really understand how to map (individual) model states to and from observational space, much less ensembles.

Parameter estimation in nonlinear models

Even with Normal input errors, nonlinearity implies non-normal output errors, complicating not only "state" estimation but also parameter selection.

"Recalibration"

Unlikely in meteorology

von Mises (1928)

Current Challenges:

Limited relevance of the Kalman Filter

"Of course, in general these tasks (prediction, separation, detection) may be done better by nonlinear filters."

(Kalman, 1960; first substantial footnote)

- Use of 4DVar with imperfect model(s)

The target is no longer a max likelihood state, in fact the model may not support the most "realistic" looking states.

- Ensemble "spread" and "bias" correction.

Distinguishing "good spread" and "bad spread" given ~ 100 points in a ~10,000,000-dim space.

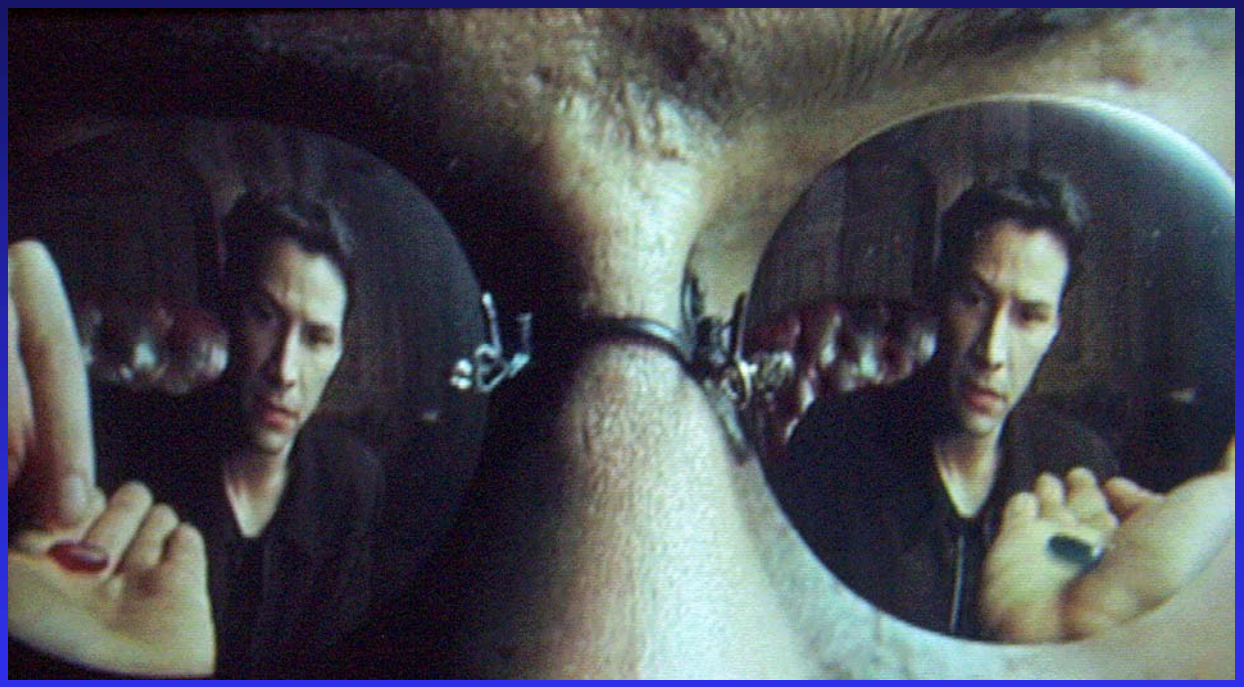
- Interpreting parametric uncertainty in the "one-off" case (climate).

What are "reasonable" parameter ranges?
How climate variables differ from weather?
Can a prior distribution and a transfer function yield a policy relevant PDF?

Applications in this Context

Model development	resource distribution for utility (not for naïve realism)
Parameter Estimation	relaxed (to within the physical relevance of then parameterisation)
Data Assimilation	allow each model its manifold, assimilate without re-simulating!
(Ensemble) Simulation	perturb as far in the past as possible: do NOT resample
Forecasting:	true eMOS
Informed Decision Making	a PDF, but not as we know it
Model(s) Improvement	evaluation & forecast archive

Now you have to make a choice. You take the blue pill and the lecture ends, you wake–up in your bed and happily do mathematics…



You take the red pill, and try to do physics/statistics in the real world knowing all models are wrong.

"Remember that all I am offering is the truth. Nothing more"
Morpheus

2 Feb 2006

LA Smith (2003) Predictability Past Predictability Present. ECMWF. soon to be in a CUP book (ed. Palmer).

LA Smith (2000) *Disentangling Uncertainty and Error*, in Nonlinear Dynamics and Statistics (ed A.Mees) Birkhauser.

K Judd and LA Smith (2001) *Indistinguishable States I* , Physica D **151**: 125-151 *(2004) Indistinguishable States II,* **196:** 224-242 .

M. Altalo and LA Smith (2004) Environmental Finance **6** (1) 48-49.

M Roulston *and LA Smith* (2003) MWR **130** (6): 1653-1660.

A Weisheimer, L.A.Smith and K Judd (2004) A New Look at DEMETER forecasts via Bounding Boxes Tellus (to appear).

LA Smith (2002) *What might we learn from climate forecasts?*, Proc. National Acad. Sci. **99**: 2487-2492.

www.lsecats.org lenny@maths.ox.ac.uk

Questions:

What are we trying to do? Exactly?
What is model inadequacy? What is "uncertainty in the initial condition"?
How to compare/combine simulations? (How do you know an improved model is better?)
How should we judge forecasts? (PDF relevant skill scores? Not RMS!)
- include the verification as on EPS member: ?what is the impact on the skill score?
- beyond a better best member (but without rejecting a perfect model!)
- and go beyond C/L (Questions of 'how much' not 'whether')

When to treat simulations as scenarios (vs product space approach)?
How to forget 'best' ? (and accept/identify a move towards better)
How to let the simulations speak for themselves?
Issues of Statistical Good Practice (the real dark side) [avoid being misled in 10^7 D]
Out-of-sample, bootstrapped significance, fair counting, Imperfect model, nonlinearity …

Things to distinguish:

Simulations from Forecasts [Forecasts can evolve after the simulations are fixed]
Probability Forecasts from Ensemble Prediction Systems [Deterministic from Unequivocal]
'Useful' spread from 'Bad spread' from 'model-optimal' spread [The path from the goal]
High Impact Forecasts from Severe WeatherForecasts
Model Variables from Physical Variables (Projection Operator P)
Empirical Adequacy vs Internal Model Consistency (Z500)
Improving tomorrow's Forecast from improving 'the' 2020 Simulation
Goal of simulations (shadowing) from that of forecasts (information on an observable)
Simulations as Scenarios from Product Space Approaches
Accurate Forecasts from Useful Forecasts (esp risk adverse users) [both sci and psych]