3rd International Verification Methods Workshop

January 29 – February 2, 2007

Ensemble Forecasts: An Introduction to Their Statistics, Value and Application

Leonard A Smith with Jochen Broecker and Hailiang Du

cats@lse.ac.uk

A general introduction to forecasting with accurate but imperfect observations and good but imperfect models is presented; the viewpoints of both the forecaster and the user(s) are considered. We'll see how in this context the idea of ensemble forecasting arises naturally: given that we do not know exactly where to start our models, it makes good sense to examine model runs from a collection of nearby initial conditions, this collection is called an ensemble. But exploiting this simple insight leads to a number of interesting statistical questions: How do we know if our ensemble forecast is "good" in this case? How do we approach questions allocating limited resources between using a more costly model, using a larger ensemble, and investing more to obtain a "better" ensemble of initial conditions in the first place? And once we have an ensemble forecast, how is a user to interpret and exploit the additional information in all these model runs? These questions are investigated in the context of medium range weather forecasting and in simple chaotic model-systems pairs where it is easier to get strong statistical evidence. A variety of tools which aid in forecast verification will be explored, including the construction of reliability diagrams and discussion of the strengths and weaknesses of various skill scores which can be applied to any probability forecast.

The discussion is motivated by and illustrated with a number of examples, some requiring real-time forecasts from the audience. While references to the technical literature will be given, the goal is to build an understanding of general concepts and how probabilistic forecasting is already changing the way industry views weather forecasts.

Ensemble Forecasts:

An Introduction to Their Statistics, Value and Application

Leonard Smith

Centre for the Analysis of Time Series, LSE

&

Pembroke College, Oxford

Jochen Broecher, Liam Clarke, Hailiang Du

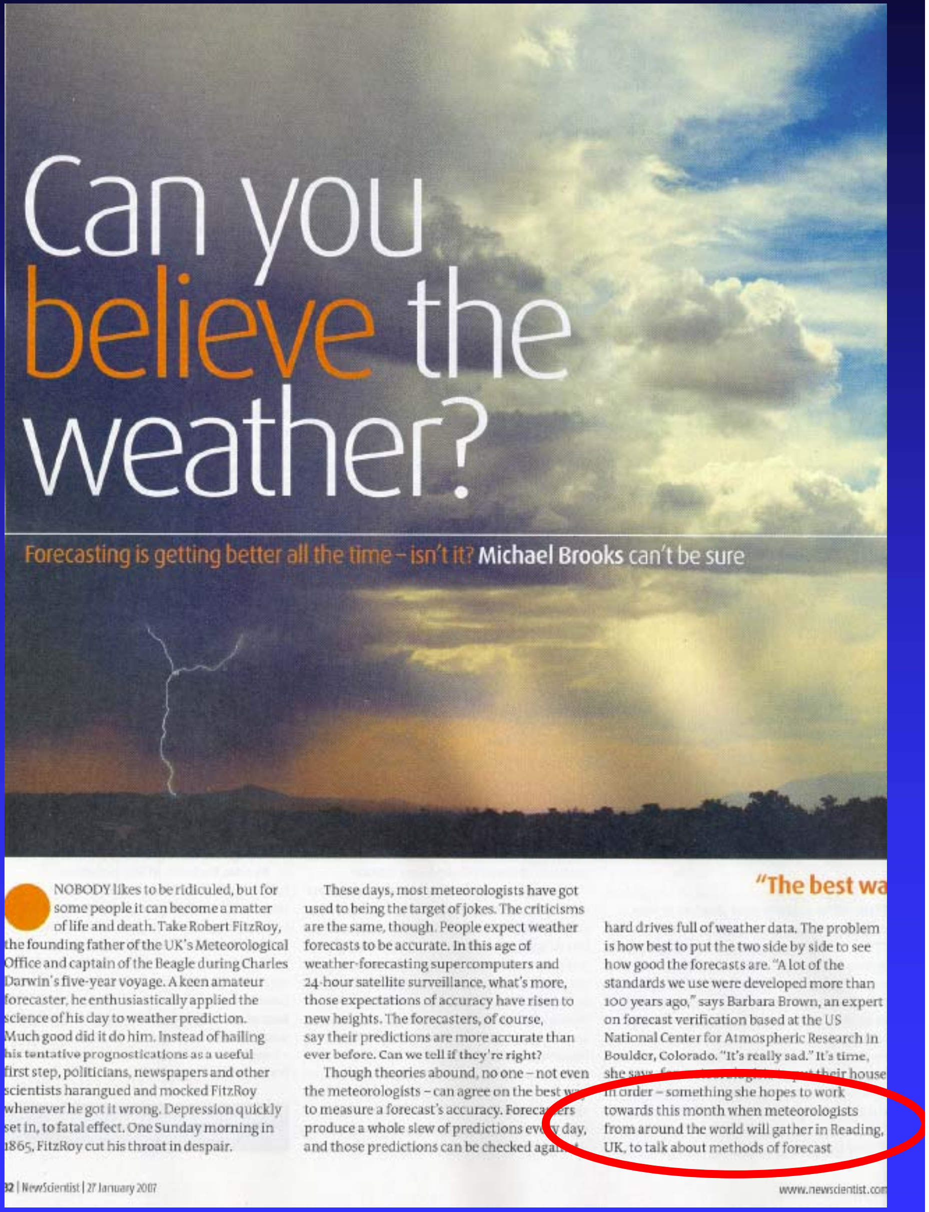
www.lsecats.org

Live Discussion Board NOW

Welcome to England.

The British have a long-standing interest in the weather.

And in science more generally…



Can you believe the weather?

Forecasting is getting better all the time – isn't it? **Michael Brooks** can't be sure

NOBODY likes to be ridiculed, but for some people it can become a matter of life and death. Take Robert FitzRoy, the founding father of the UK's Meteorological Office and captain of the Beagle during Charles Darwin's five-year voyage. A keen amateur forecaster, he enthusiastically applied the science of his day to weather prediction. Much good did it do him. Instead of hailing his tentative prognostications as a useful first step, politicians, newspapers and other scientists harangued and mocked FitzRoy whenever he got it wrong. Depression quickly set in, to fatal effect. One Sunday morning in 1865, FitzRoy cut his throat in despair.

These days, most meteorologists have got used to being the target of jokes. The criticisms are the same, though. People expect weather forecasts to be accurate. In this age of weather-forecasting supercomputers and 24-hour satellite surveillance, what's more, those expectations of accuracy have risen to new heights. The forecasters, of course, say their predictions are more accurate than ever before. Can we tell if they're right?

Though theories abound, no one – not even the meteorologists – can agree on the best way to measure a forecast's accuracy. Forecasters produce a whole slew of predictions every day, and those predictions can be checked against

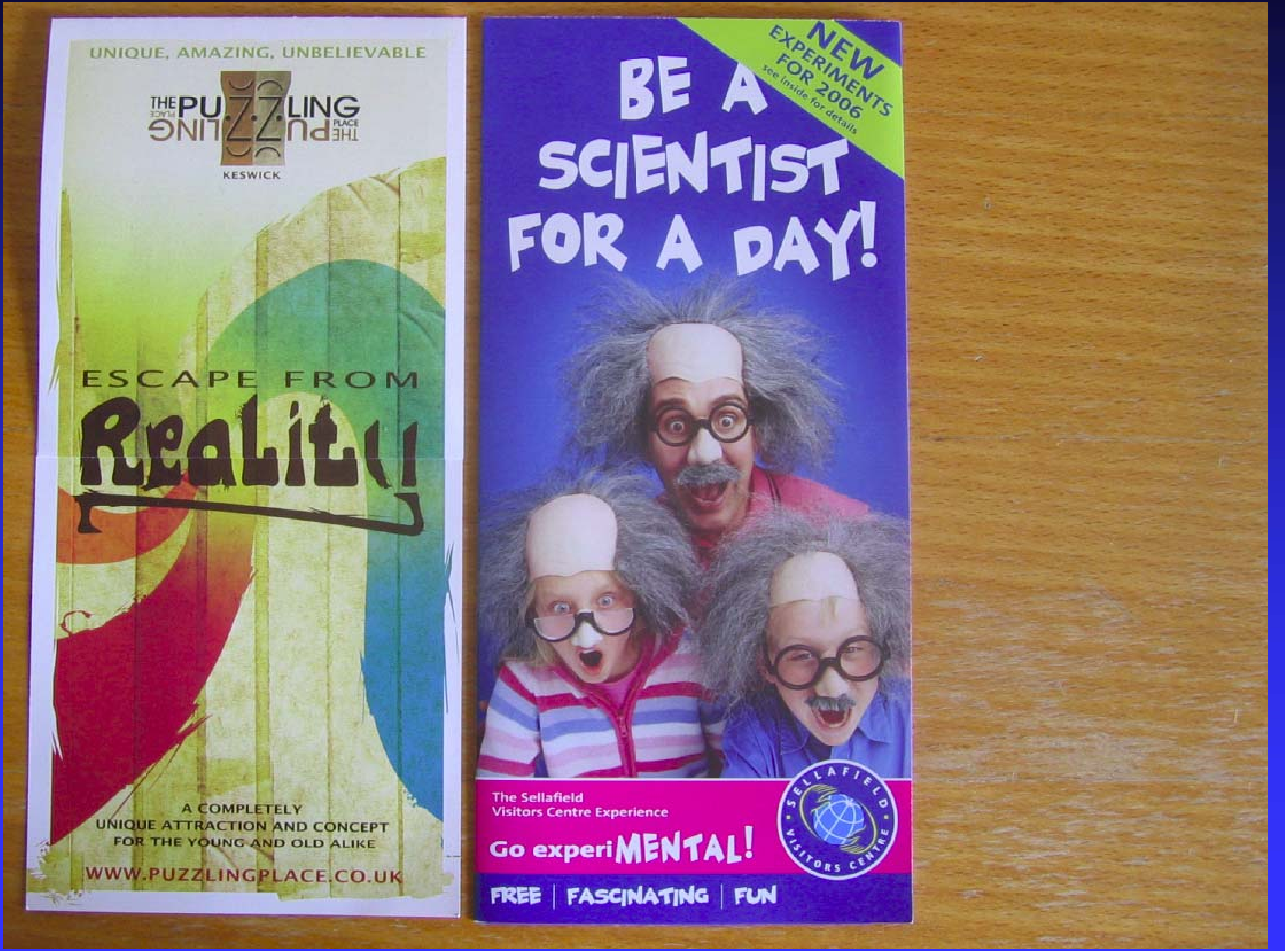
"The best wa

hard drives full of weather data. The problem is how best to put the two side by side to see how good the forecasts are. "A lot of the standards we use were developed more than 100 years ago," says Barbara Brown, an expert on forecast verification based at the US National Center for Atmospheric Research in Boulder, Colorado. "It's really sad." It's time, she says, for meteorologists to put their house in order – something she hopes to work towards this month when meteorologists from around the world will gather in Reading, UK, to talk about methods of forecast

32 | NewScientist | 27 January 2007

www.newscientist.com

Welcome to England.


UNIQUE, AMAZING, UNBELIEVABLE
THE PUZZLING PLACE
KESWICK
ESCAPE FROM
Reality
A COMPLETELY UNIQUE ATTRACTION AND CONCEPT FOR THE YOUNG AND OLD ALIKE
WWW.PUZZLINGPLACE.CO.UK
NEW EXPERIMENTS FOR 2006
see inside for details
BE A SCIENTIST FOR A DAY!
The Sellafield Visitors Centre Experience
Go experiMENTAL!
SELLAFIELD VISITORS CENTRE
FREE | FASCINATING | FUN

Welcome to England.

learnt about gravity will be turned on its head. Can a ball roll uphill? Will water really flow uphill? Only when you've experienced it will you believe it!
...and MUCH MUCH MORE ...you NEED to see for YOURSELF!
THE EXHIBITION
Many of the best optical illusions around are demonstrated through interactive exhibits, artwork and sculpture. You'll be left, quite literally, not believing your eyes!
YES
INTO A WORLD OF
ILLUSION
A COMPLETELY UNIQUE ATTRACTION AND CONCEPT FOR THE YOUNG AND OLD ALIKE
WWW.PUZZLINGPLACE.CO.UK
NEW EXPERIMENTS FOR 2006
see inside for details
BE A SCIENTIST FOR A DAY!
The Sellafield Visitors Centre Experience
Go experiMENTAL!
SELLAFIELD VISITORS CENTRE
FREE | FASCINATING | FUN

Although the phrase 'forecast verification' is generally used in atmospheric sciences, it is rarely used outside the discipline.

Jolliffe and Stephenson, pg *xi*

Verification & Validation ↔ Evaluation, Tuning & Value

Verification Methods can:

- help us compare and evaluate different forecasts
- help us tune our forecast systems
- help us provide higher value forecasts to numerate users
- help us improve our models and our entire EPS system coherently
- help users combine and exploit every forecast worth paying for
- and help users decide how many to buy and how much to pay.

In short: in terms of forecast performance they help us distinguish reality & illusion.

3rd International Verification Methods Workshop
January 29 – February 2, 2007
To be held at ECMWF, Reading, UK
Click here to find out more!

WMO Verification Workshop @ ECMWF

Three very different aims of “Verification”:

Strategic	Tactical	Rest & Recreation
Towards better models tomorrow	Better decisions given today’s state-of-the-art models.	Interesting applied maths and statistics

Verification helps us distinguish reality (skill) from illusion (luck); and verification plays different roles in EPS development, EPS use, and maths.

(and, of course, all models are wrong)

Mantra: Uncertainty, inadequacy and verification, value

Overview of the Morning

Introduction: What is “verification” and why bother?
Comments on the techniques presented earlier
Designer Verification Scores
Ensemble forecasting
Forecasting the NAG Board for insight, and an ECMWF pen!

Coffee

Quantifying Skill: Scorology
Weather Roulette
Evaluation of *your* NAG Board Forecasts: ECMWF pen
Users, meteorological skill scores & useful ensemble forecasting

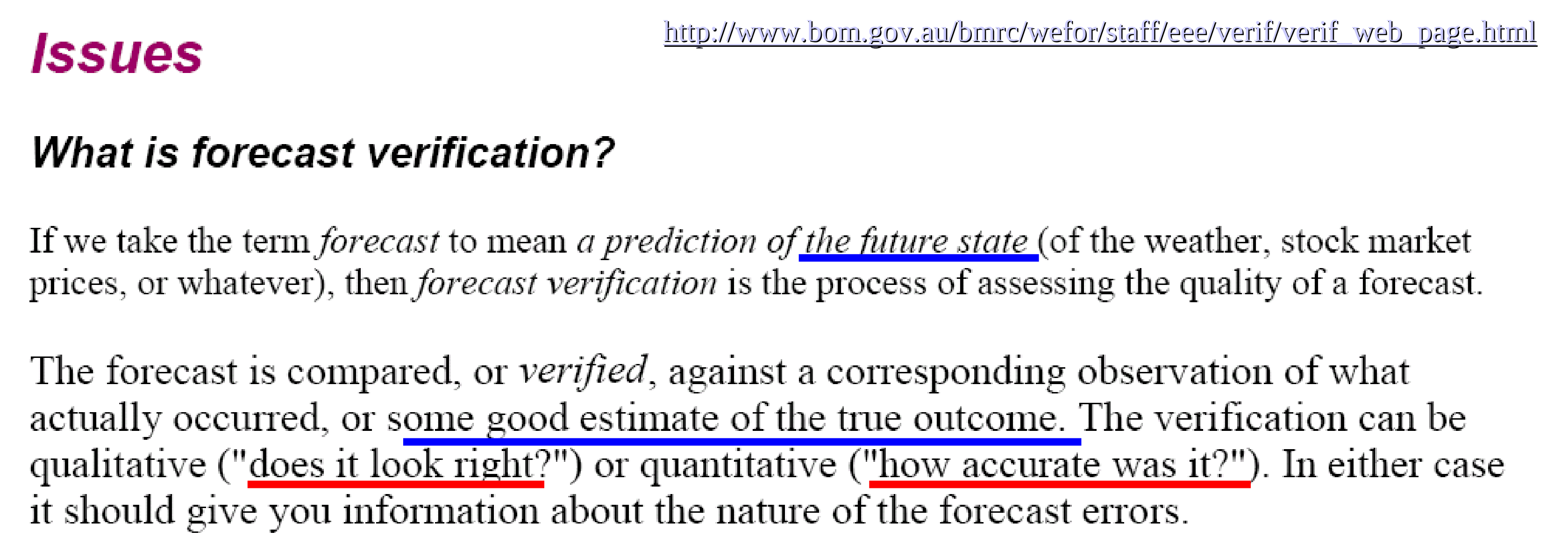
Lunch

One nice thing about giving the last tutorial is that most of the definitions have already been given, and we can look at “when” “what” and “why” rather than “how.”

So lets look at a few examples;

Rank Histograms (in one dimension and 10^7),
Ensemble Estrangement graphs,
Reliability Diagrams,
And a skill score (ignorance).

But first: what is verification?



http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Issues

What is forecast verification?

If we take the term *forecast* to mean *a prediction of the future state* (of the weather, stock market prices, or whatever), then *forecast verification* is the process of assessing the quality of a forecast.

The forecast is compared, or *verified*, against a corresponding observation of what actually occurred, or some good estimate of the true outcome. The verification can be qualitative ("does it look right?") or quantitative ("how accurate was it?"). In either case it should give you information about the nature of the forecast errors.

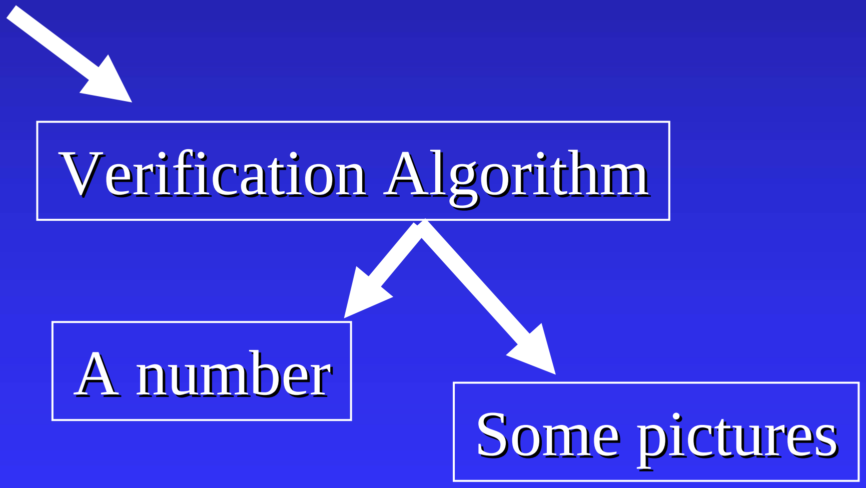
But we do not predict "the" future state we predict a distribution;
And we never learn the future state, but only (noisy, partial) observations…

So how might a probabilistic forecast "look right"? Or "be accurate"?
And what kind of information could inform us as to the nature of the forecast errors, when every forecast is a distribution, and every observation is at best a number (or vector)?

Even if we only run the model once, we are making probabilistic forecasts.

Verification can only evaluate the *entire* Ensemble Prediction System

Observational noise model
Data assimilation scheme
Ensemble formation scheme
Forecast model
Ensemble interpretation scheme
→ probabilistic forecast



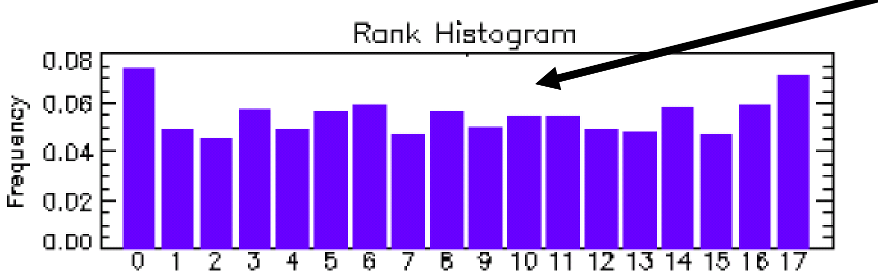
Verification Algorithm

A number

Some pictures

What you do with that number or those diagrams will depend on why you are looking at the forecast in the first place.

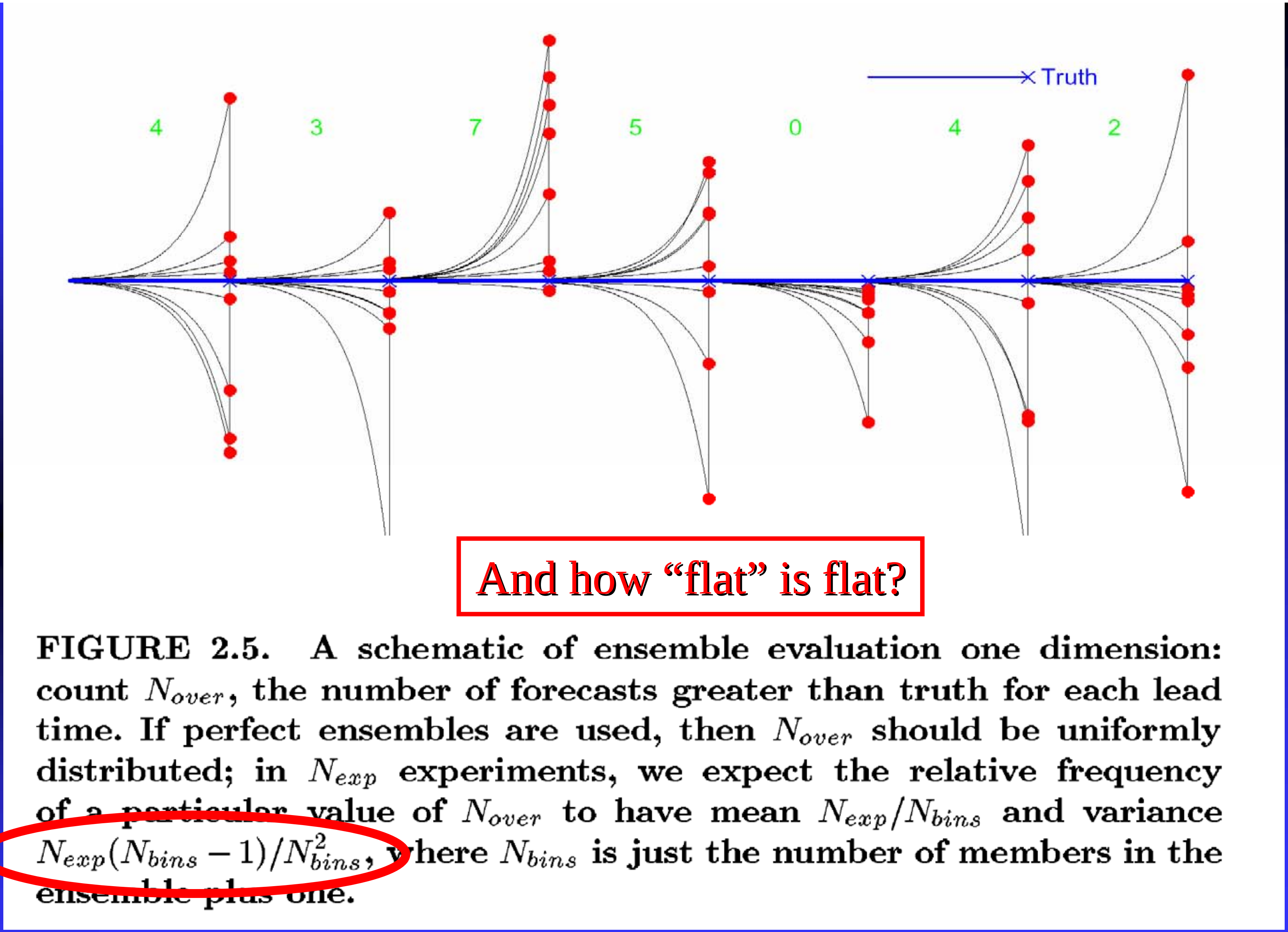
Rank histogram (Talagrand et al, 1997; Hamill, 2001)



And is that “flat”?
Care to vote?

Answers the question: *How well does the ensemble spread of the forecast represent the true variability (uncertainty) of the observations?*

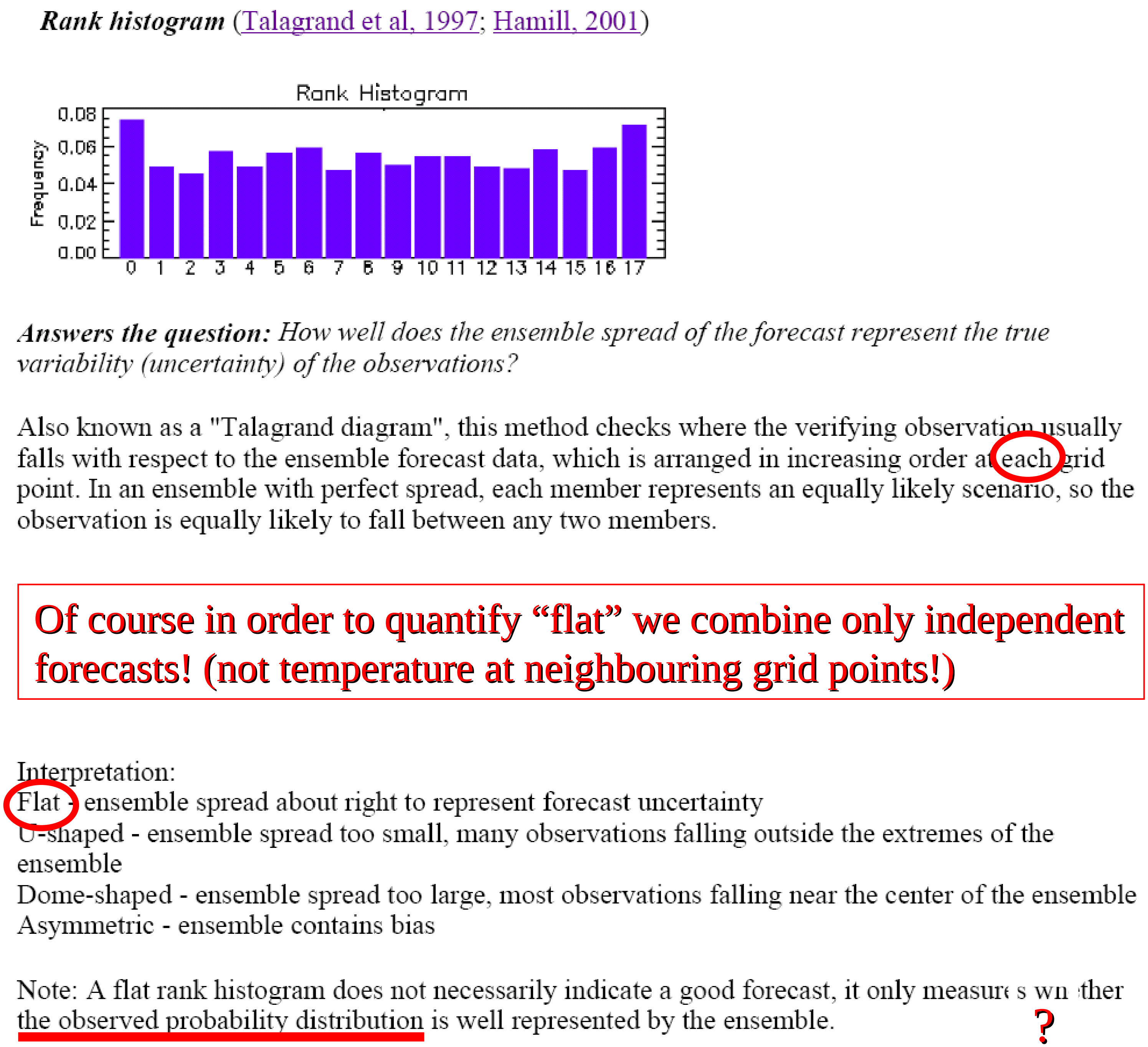
Also known as a "Talagrand diagram", this method checks where the verifying observation usually falls with respect to the ensemble forecast data, which is arranged in increasing order at each grid point. In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members.



And how “flat” is flat?

FIGURE 2.5. A schematic of ensemble evaluation one dimension: count N_{over}, the number of forecasts greater than truth for each lead time. If perfect ensembles are used, then N_{over} should be uniformly distributed; in N_{exp} experiments, we expect the relative frequency of a particular value of N_{over} to have mean N_{exp}/N_{bins} and variance $N_{exp}(N_{bins}-1)/N_{bins}^2$, where N_{bins} is just the number of members in the ensemble plus one.

LA Smith (2000) 'Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems' in *Nonlinear Dynamics and Statistics*, ed. Alistair I. Mees, Boston: Birkhauser, 31-64



Rank histogram (Talagrand et al, 1997; Hamill, 2001)

Answers the question: *How well does the ensemble spread of the forecast represent the true variability (uncertainty) of the observations?*

Also known as a "Talagrand diagram", this method checks where the verifying observation usually falls with respect to the ensemble forecast data, which is arranged in increasing order at each grid point. In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members.

Of course in order to quantify "flat" we combine only independent forecasts! (not temperature at neighbouring grid points!)

Interpretation:
Flat - ensemble spread about right to represent forecast uncertainty
U-shaped - ensemble spread too small, many observations falling outside the extremes of the ensemble
Dome-shaped - ensemble spread too large, most observations falling near the center of the ensemble
Asymmetric - ensemble contains bias

Note: A flat rank histogram does not necessarily indicate a good forecast, it only measures whether the observed probability distribution is well represented by the ensemble.

?

To use rank histograms in higher dimensions, as with a gridded temperature field, we need to introduce "minimum spanning trees"; but the idea is the same as in 1-D.

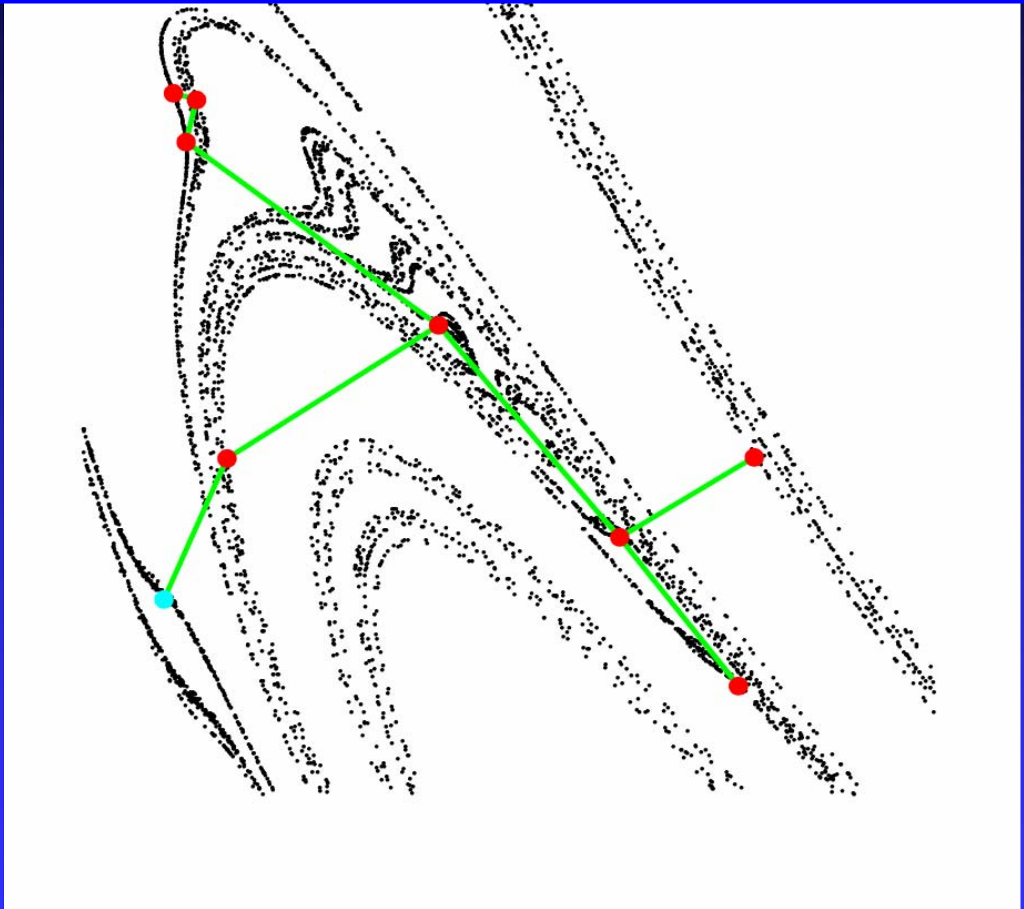


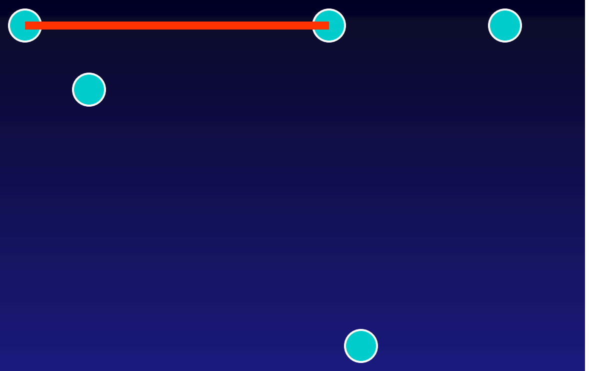
FIGURE 2.6. A minimal spanning tree from the combined set of 8 ensemble members (dark dots) and the verification (light dot) which is also on the attractor (and in this experiment "truth').

See also:

LA Smith & JA Hansen (2004) *Mon. Weather Rev.* 132 (6): 1522-1528

Wilks, DS (2004) *Mon. Weather Rev.* 132: 1329-1340.

Definitions



Given a set of points:

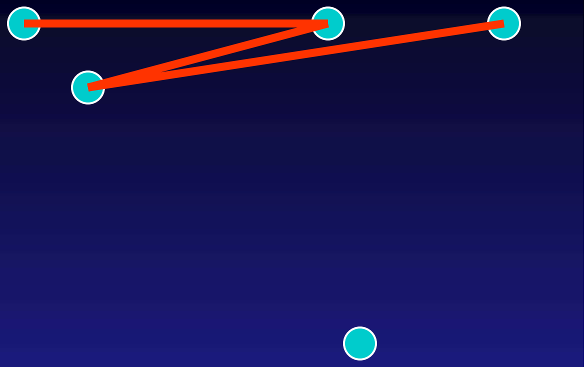
A branch is a line that connects two points

A tree is a collection of branches

A spanning tree is a collection of branches that connects all points

A minimum spanning tree is the spanning tree with the smallest sum of branch lengths

Definitions



Given a set of points:

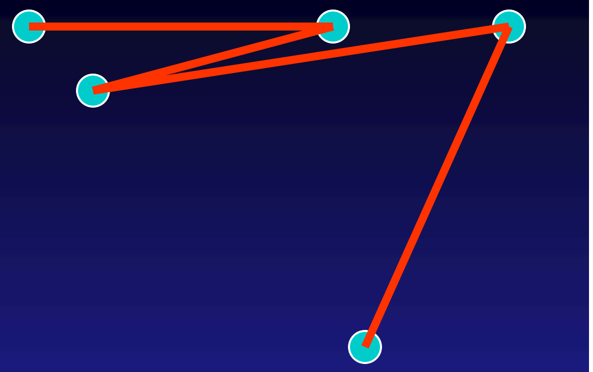
A branch is a line that connects two points

A tree is a collection of branches

A spanning tree is a collection of branches that connects all points

A minimum spanning tree is the spanning tree with the smallest sum of branch lengths

Definitions



Given a set of points:

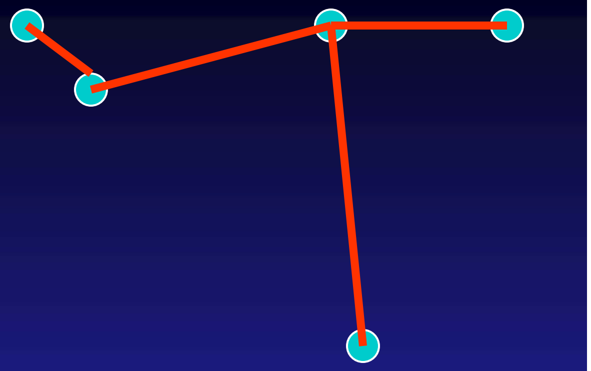
A branch is a line that connects two points

A tree is a collection of branches

A spanning tree is a collection of branches that connects all points

A minimum spanning tree is the spanning tree with the smallest sum of branch lengths

Definitions



Given a set of points:

A branch is a line that connects two points

A tree is a collection of branches

A spanning tree is a collection of branches that connects all points

A minimum spanning tree is the spanning tree with the smallest sum of branch lengths

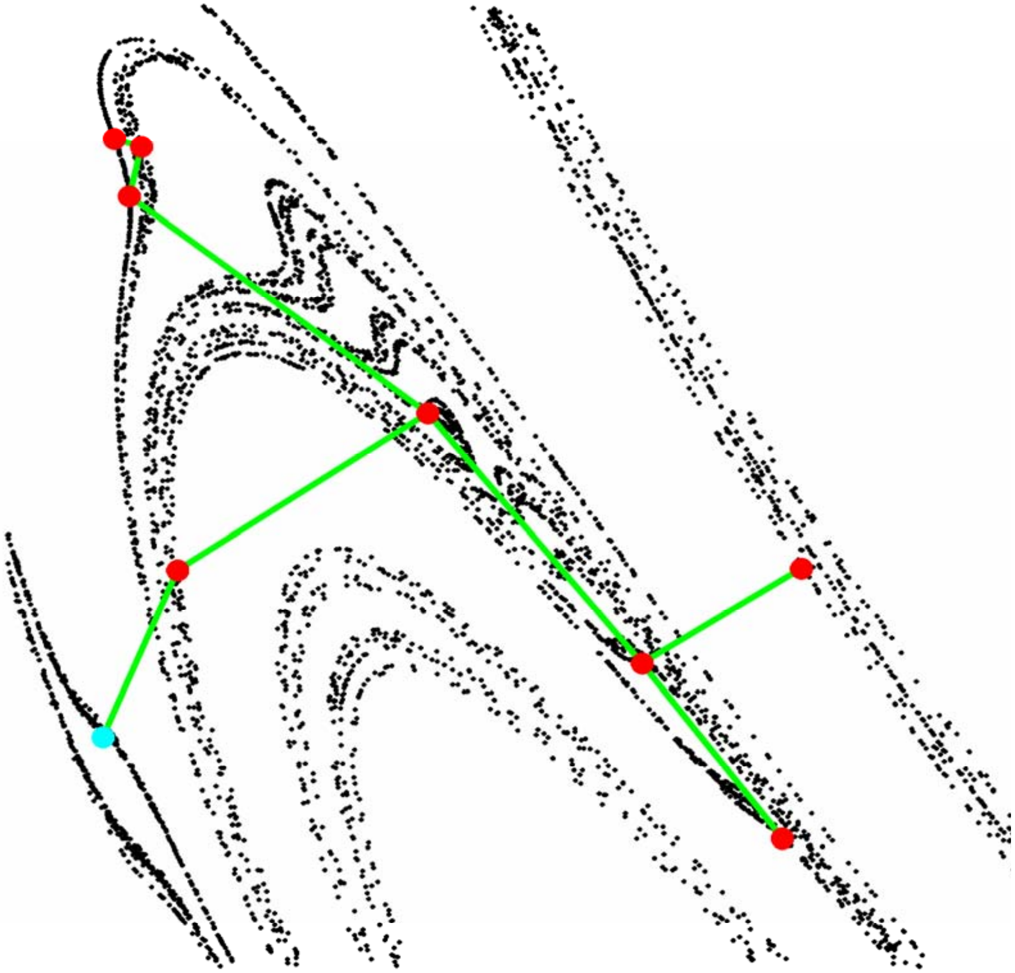
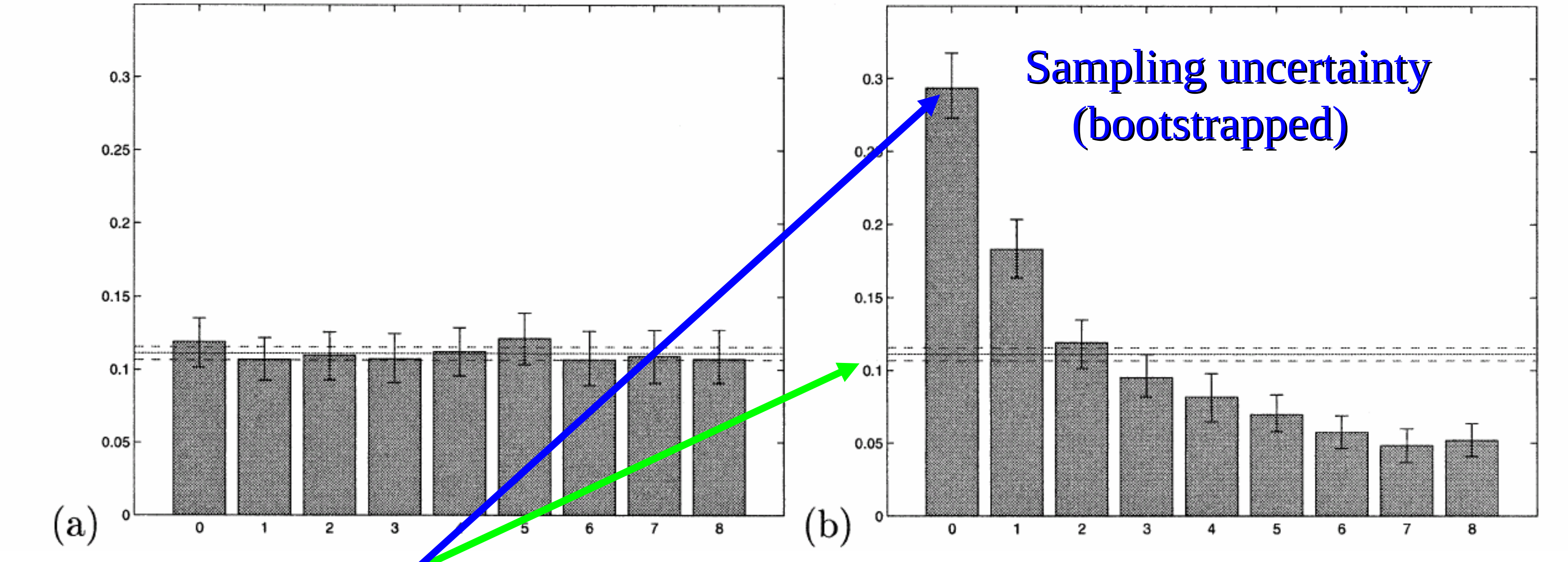


FIGURE 2.6. A minimal spanning tree from the combined set of 8 ensemble members (dark dots) and the verification (light dot) which is also on the attractor (and in this experiment "truth').

Compute all drop-one-out trees: where does the length of the tree where we dropped out the target fall in this 1-D ordering?

LA Smith & JA Hansen (2004) Extending the Limits of Forecast Verification with the Minimum Spanning Tree, *Mon. Weather Rev.* 132 (6): 1522-1528



Sampling uncertainty (bootstrapped)

Note that there are two uncertainty ranges shown on these graphs

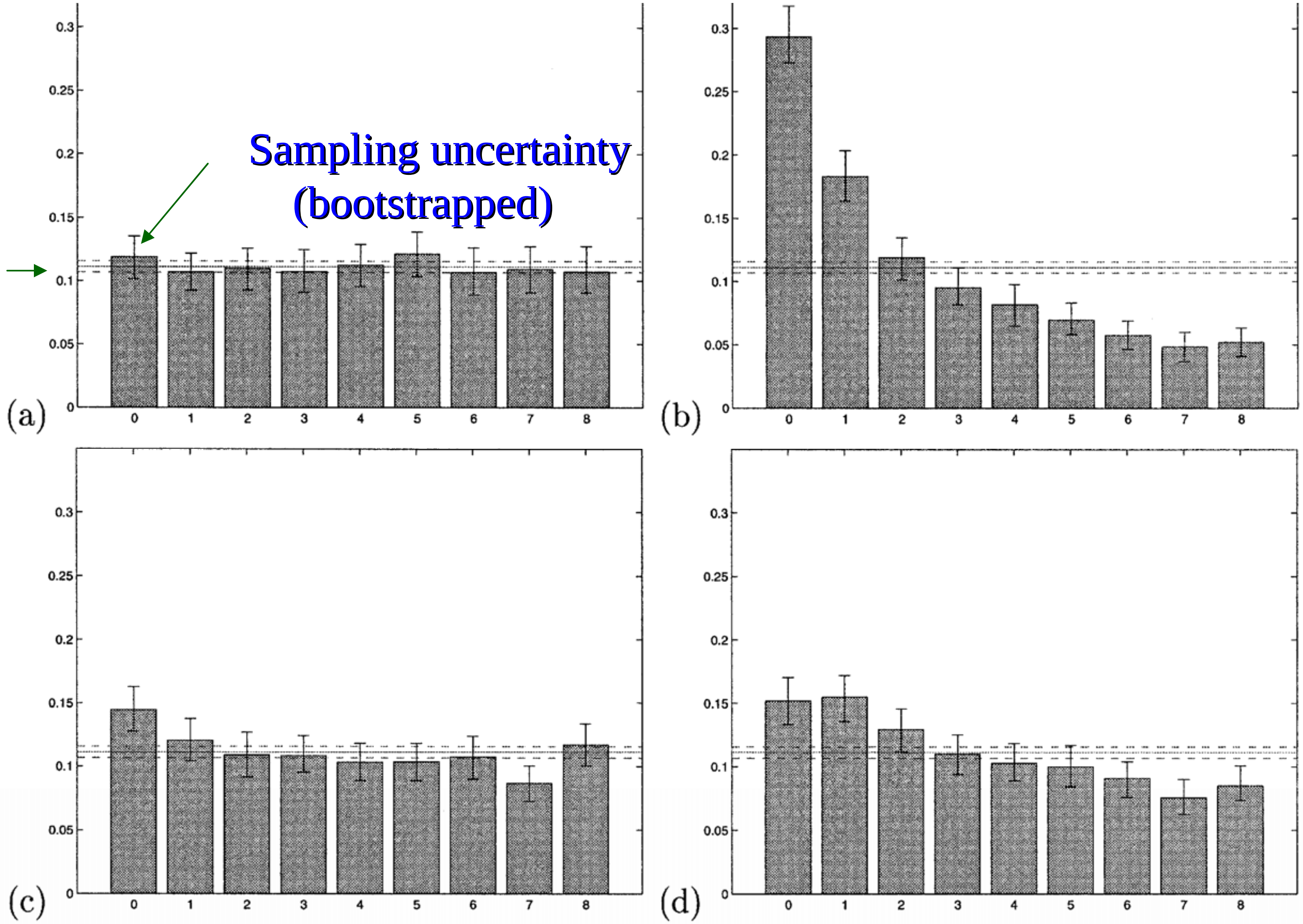
One denotes the range of variation acceptable under our null hypothesis that the ensemble distribution reflects "that of the truth."

The other denotes the likely variation in our result due to having a limited number of forecast verification pairs.

The main value here is as a consistency check; it is rarely easy to say one EPS is better than another if neither "pass" the test.

FIG. 2. Minimum spanning tree rank histograms. Ensemble members are always drawn from the Ikeda attractor, while verification differs for each panel: (a) verification is also drawn from the attractor, (b) verification is drawn randomly from a box in the area of interest, (c)

These graphs show the strength of the MST, but not its relevance!



Sampling uncertainty (bootstrapped)

Consistency

FIG. 2. Minimum spanning tree rank histograms. Ensemble members are always drawn from the Ikeda attractor, while verification differs for each panel: (a) verification is also drawn from the attractor, (b) verification is drawn randomly from a box in the area of interest, (c) verification is drawn randomly from a line that is the best linear fit to the local Ikeda attractor structure, and (d) the x component and the y component of the verification are drawn *independently* from the Ikeda attractor distribution. The solid horizontal line is the expected mean, and the horizontal dashed lines are the expected 1 std dev bounds. The vertical lines at the top of the bar in each bin are produced by bootstrapping (resampling with replacement) from the data that was used to construct the rank histograms. They represent the 99% bound on expected values.

What is the null that rank histograms test?

That the target was drawn from the same distribution that the ensemble was drawn from.

That is a pretty tough test to expect to pass!

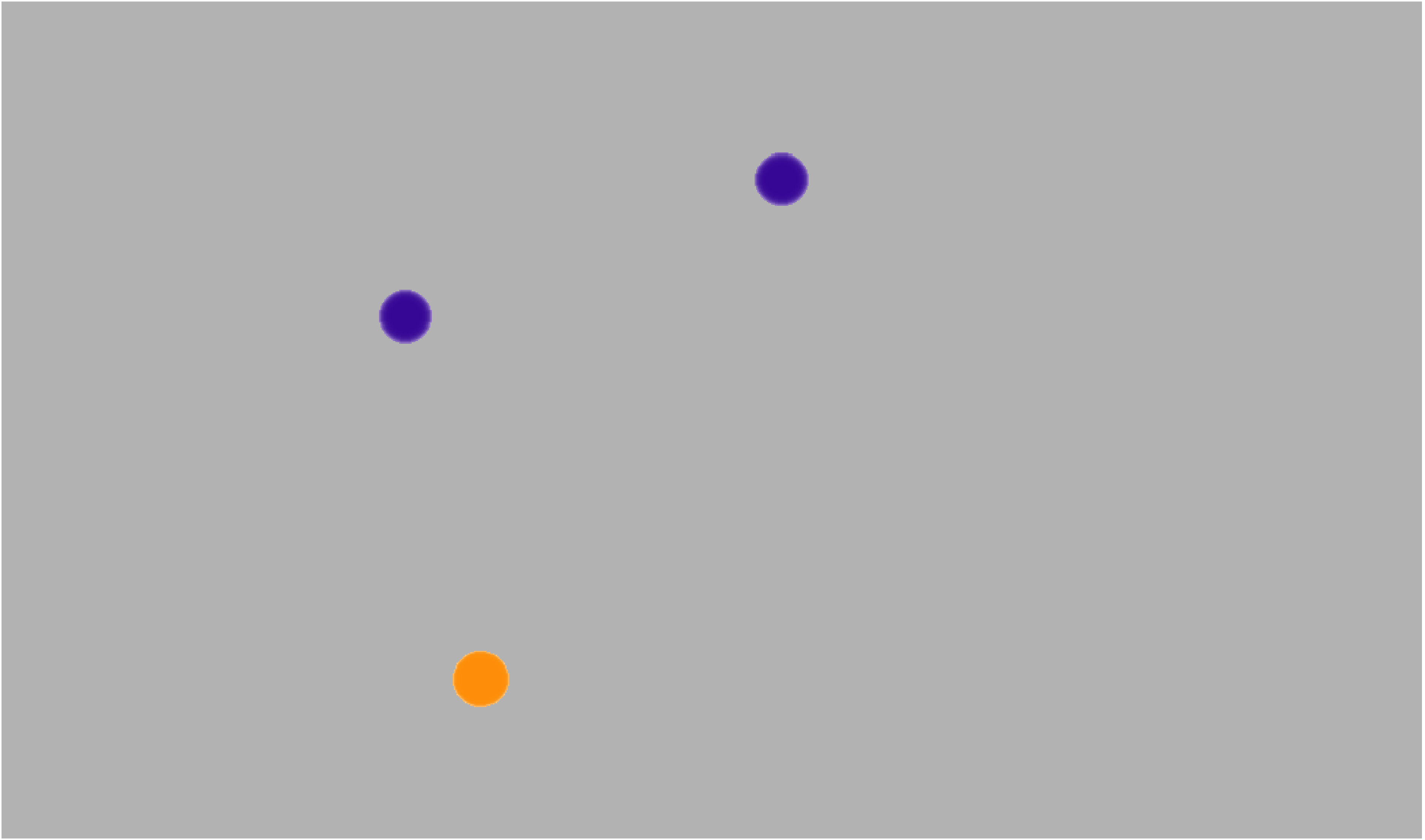
Sometimes is it more useful to make up a new test, in order to test something relevant to you; for example:

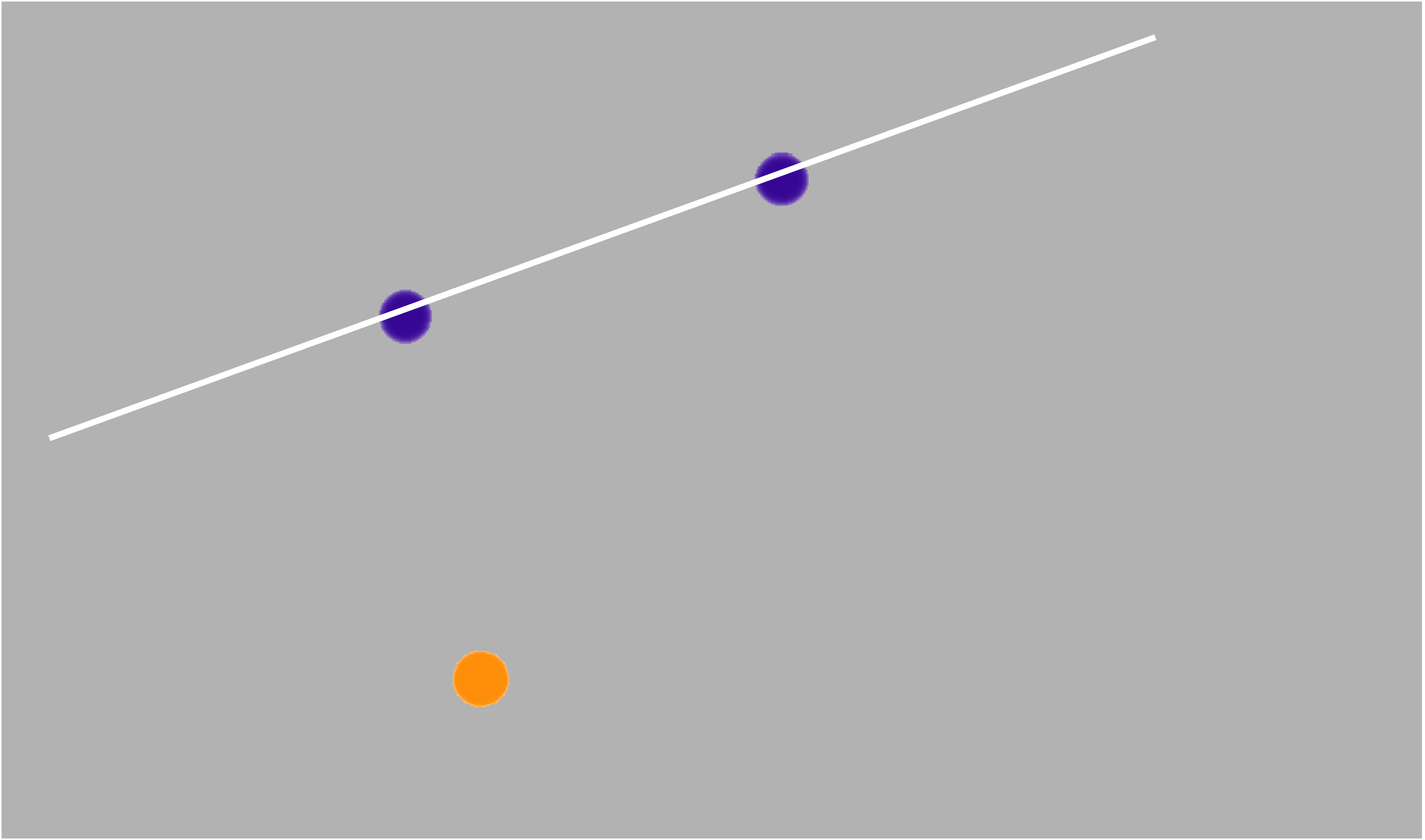
Ensemble Estrangement

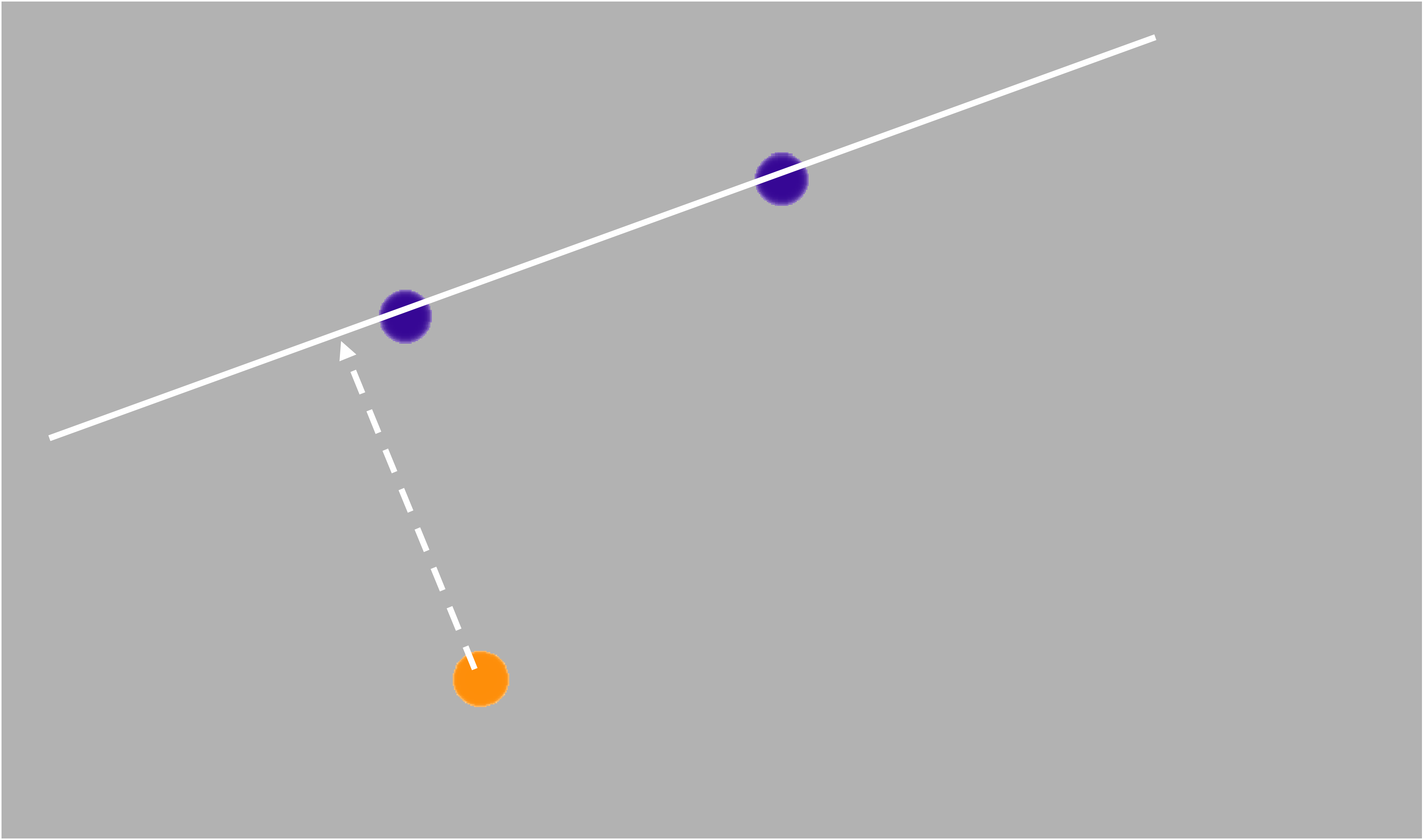
Suppose I think my ensemble has too small a spread (in a high dimensional space) and we want to "inflate" it: increase the variance of the ensemble in order to "capture" the target.

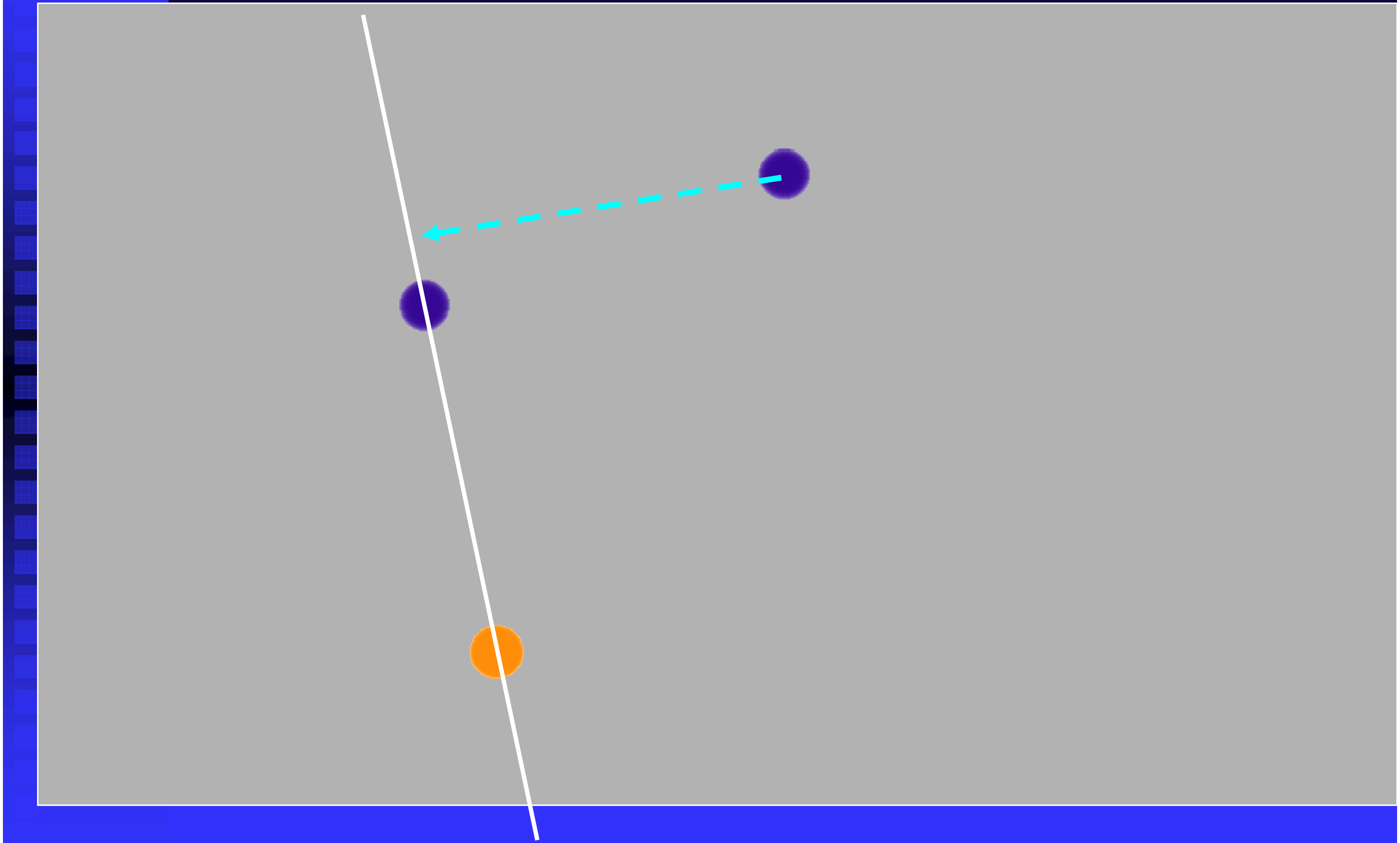
I can verify that this is a problem by testing another drop one out symmetry between the ensemble members and the target…

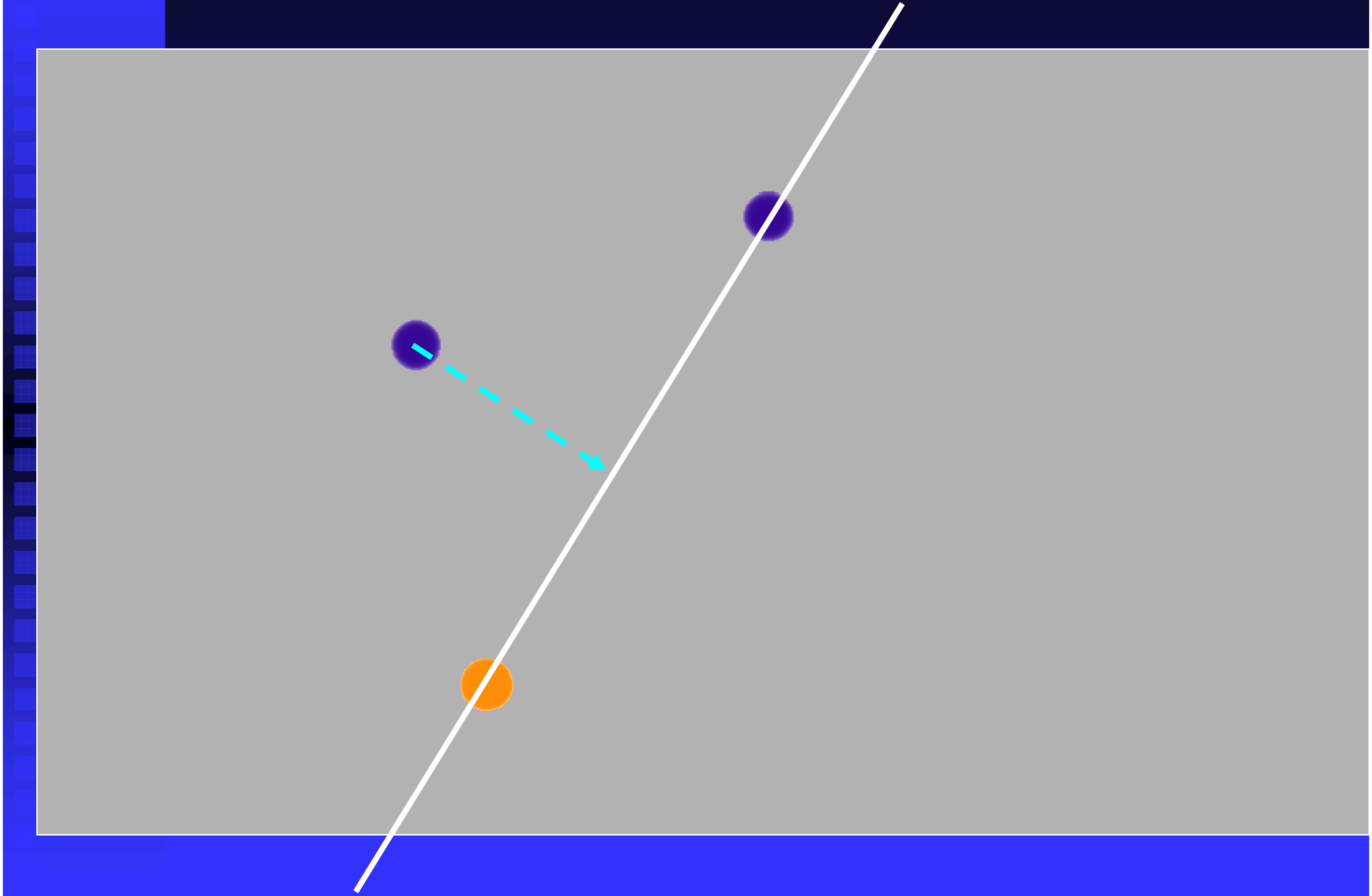
Any 51 points in a 10^7 space will lie in the same ‘line’.



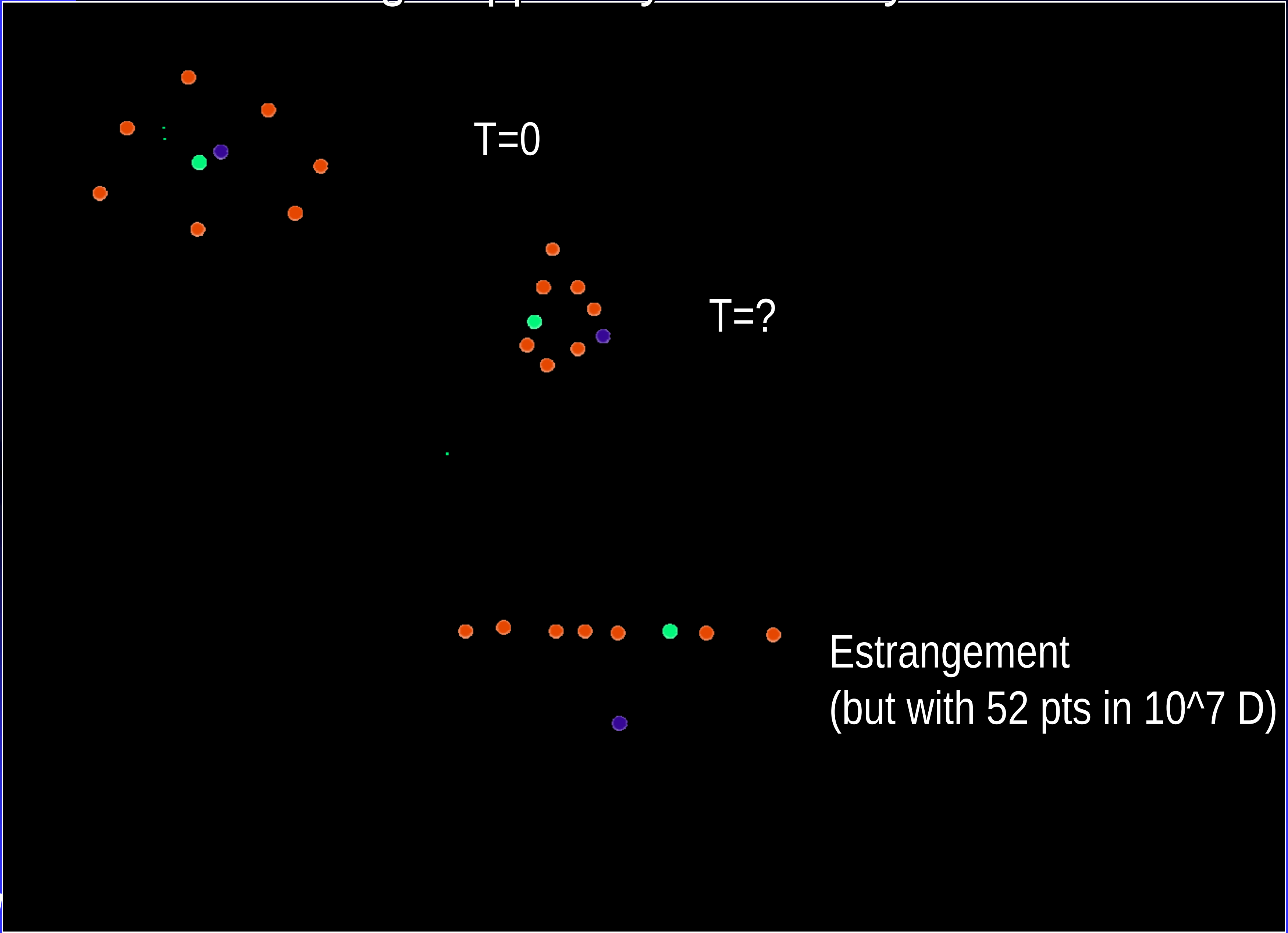








Does the target appear systematically different?

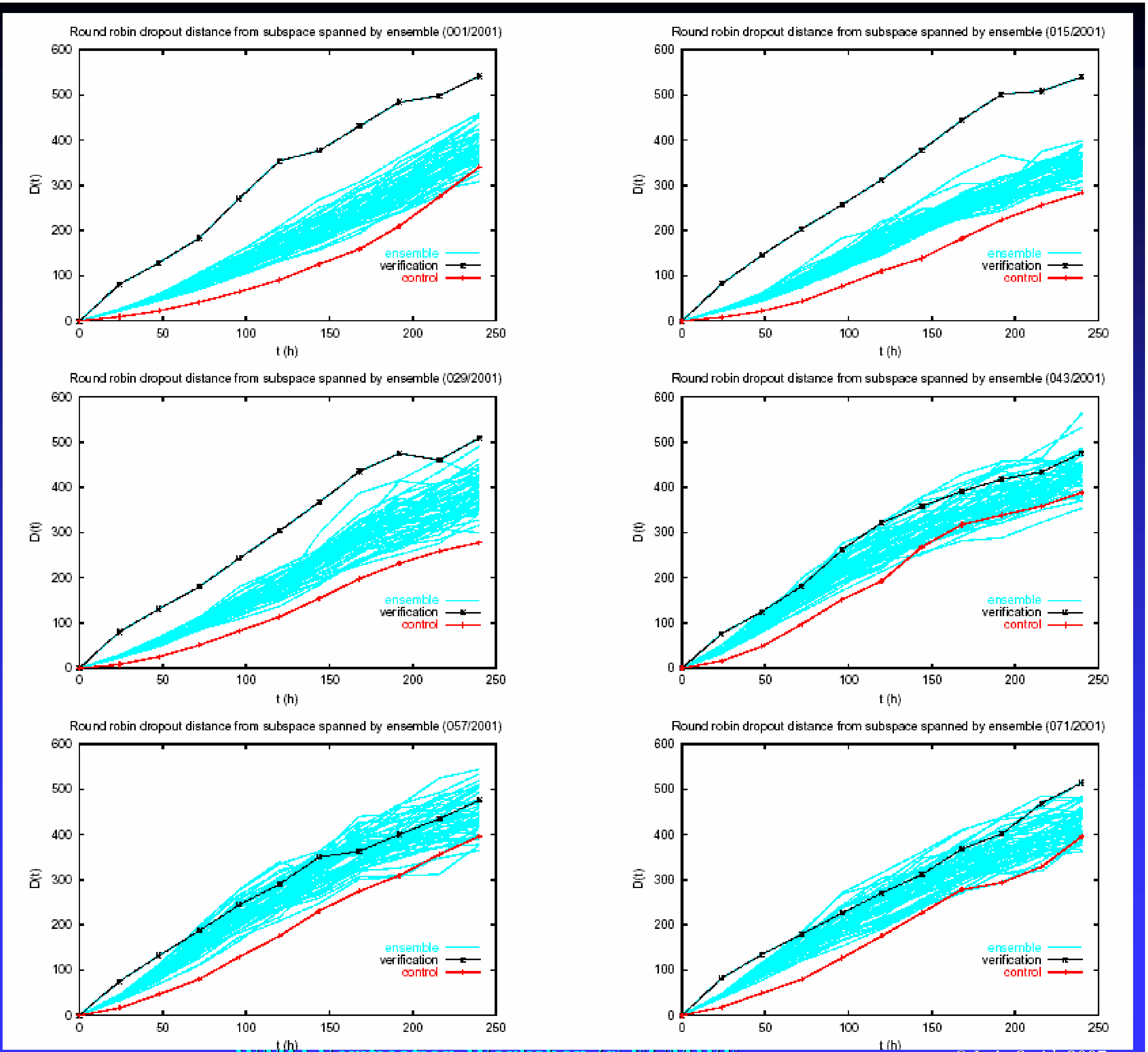


T=0

T=?

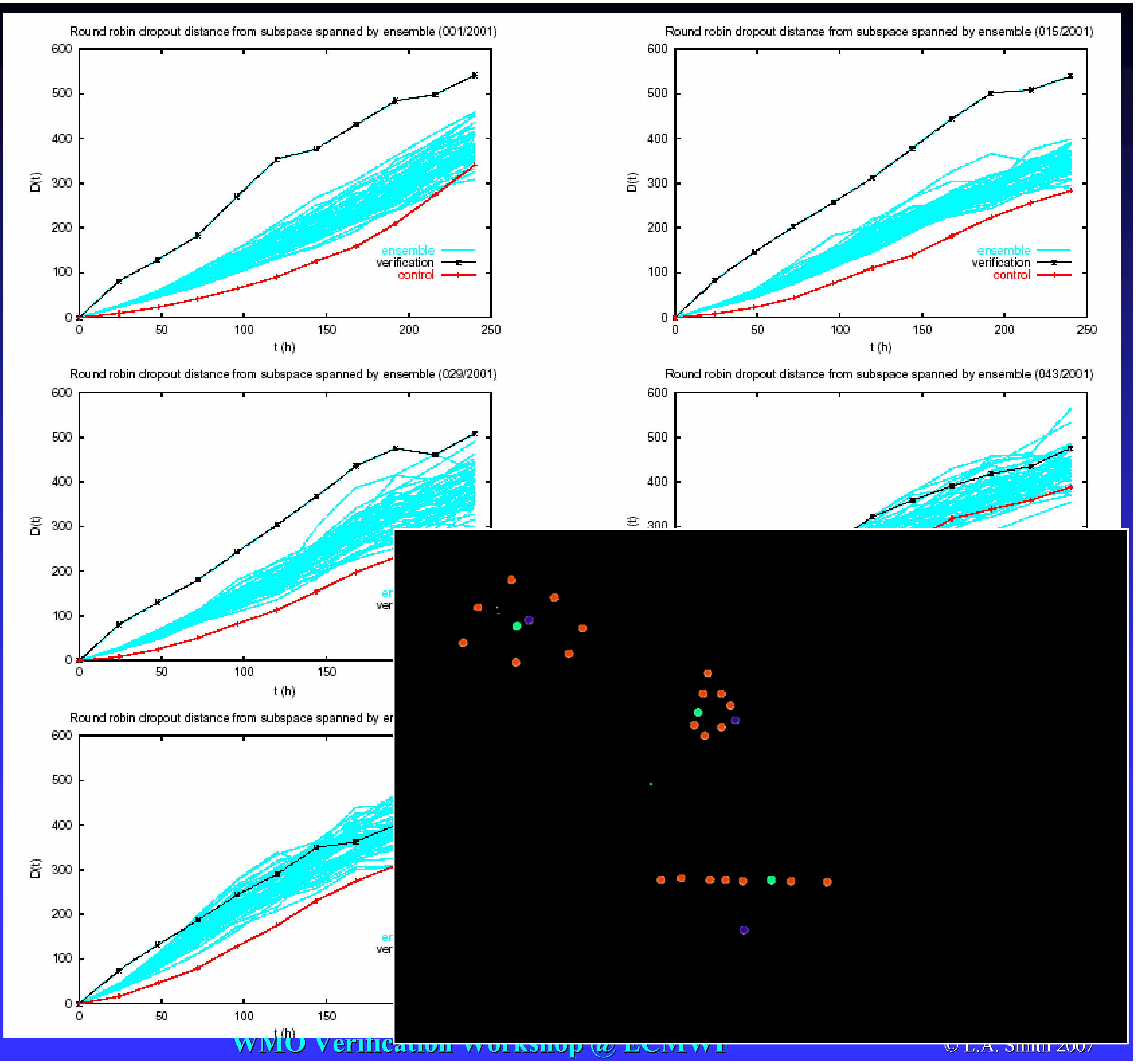
Estrangement
(but with 52 pts in 10^7 D)

Bad Spread (with the correct magnitude)

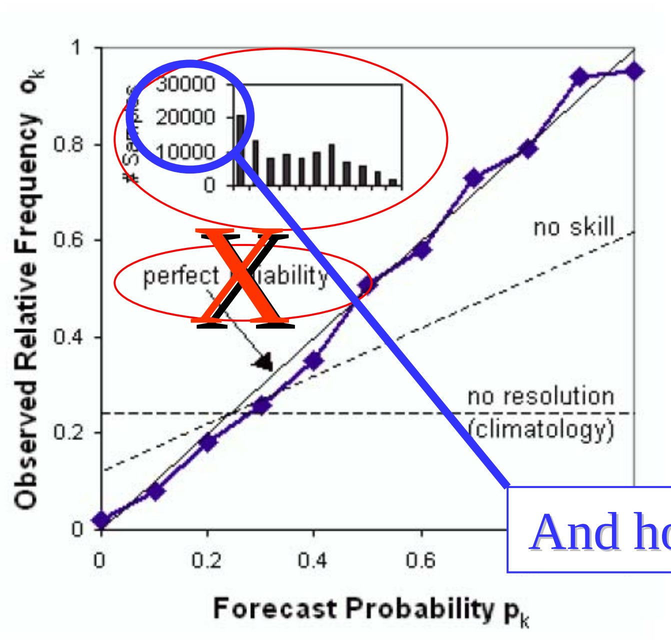


Bad Spread (with the correct magnitude!)

[Inflation will not help here!]



Reliability diagram - (called "attributes diagram" when the no-resoloution and no-skill w.r.t. climatology lines are included).



Even perfectly reliable forecasts will not lie upon the diagonal, and visually it is difficult to determine when they are “close” to it; Jochen will discuss ways forward here after the break.

And how did we get 70000 independent forecasts!

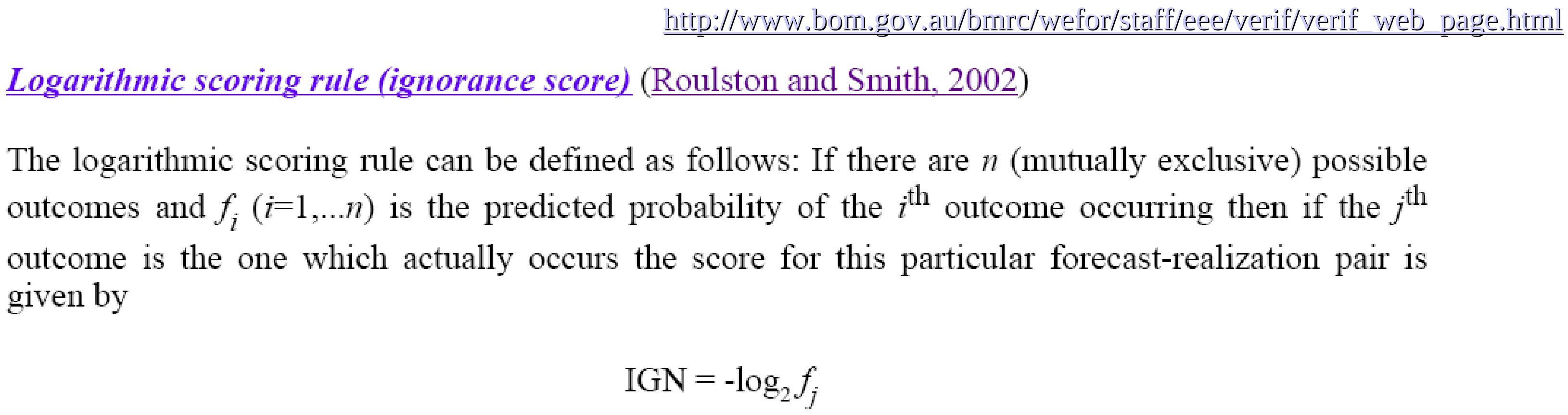
The reliability diagram plots the observed frequency against the forecast probability, where the range of forecast probabilities is divided into *K* bins (for example, 0-5%, 5-15%, 15-25%, etc.). The sample size in each bin is often included as a histogram or values beside the data points.

Answers the question: *How well do the predicted probabilities of an event correspond to their observed frequencies?*

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

In fact, as presented here it answers the question are the predicted probabilities equal to the observed relative frequencies.

And while it is clear how to draw a reliability diagram for prob(freezing), how do we use this for PDF forecasts of the noon temperature at LHR?



http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Logarithmic scoring rule (ignorance score) (Roulston and Smith, 2002)

The logarithmic scoring rule can be defined as follows: If there are n (mutually exclusive) possible outcomes and f_i (i=1,...n) is the predicted probability of the i^{th} outcome occurring then if the j^{th} outcome is the one which actually occurs the score for this particular forecast-realization pair is given by

$$\mathrm{IGN} = -\log_2 f_j$$

Good, I.J., 1952: Rational decisions, *J. Royal Statistical Soc*, **14**, 107-114

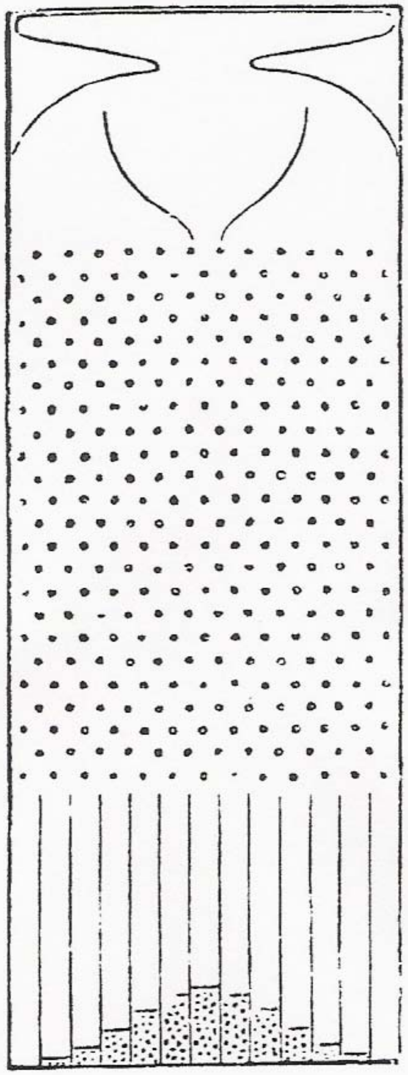
Good (1952) who suggested that the met office should be rewarded for improvements in this value.

It is effectively unique, as it is proper and local. (Jochen)

And it is useful even when f_j is not a probability…

But what do you do, given an ensemble?

This Galton Board is a mathematical model.



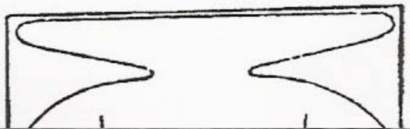
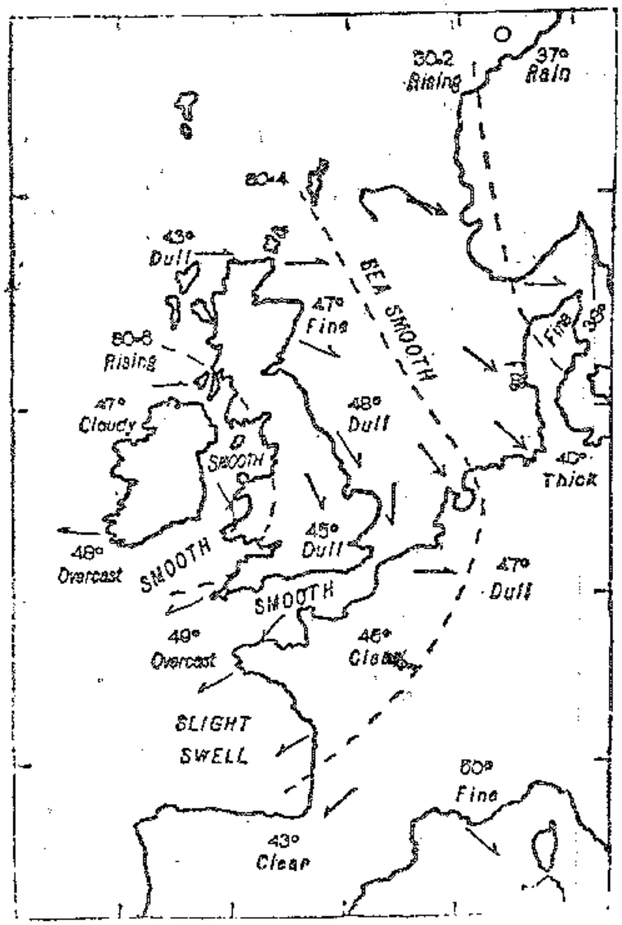
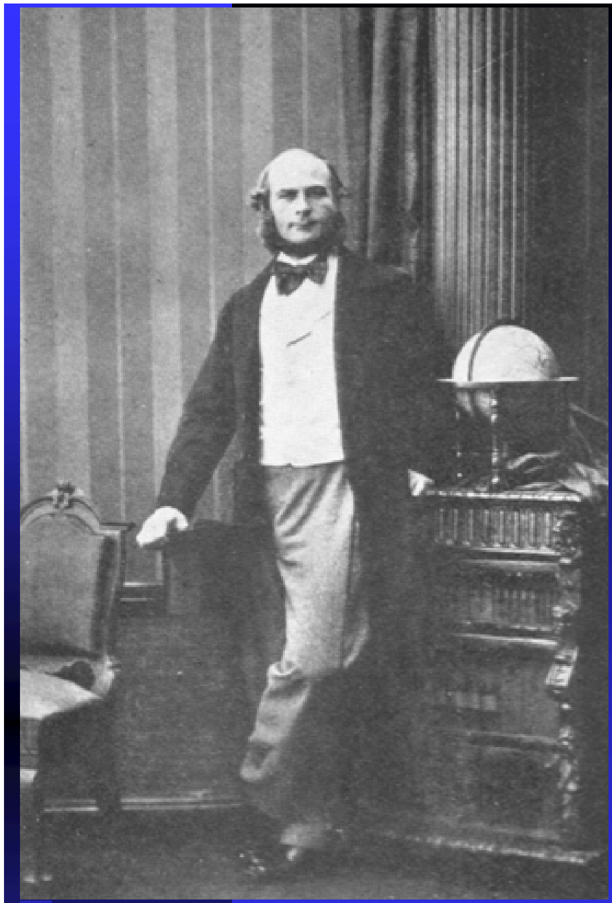


Figure 9.2 A schematic drawing of Galton's Quincunx, from Galton (1889a, p. 63).

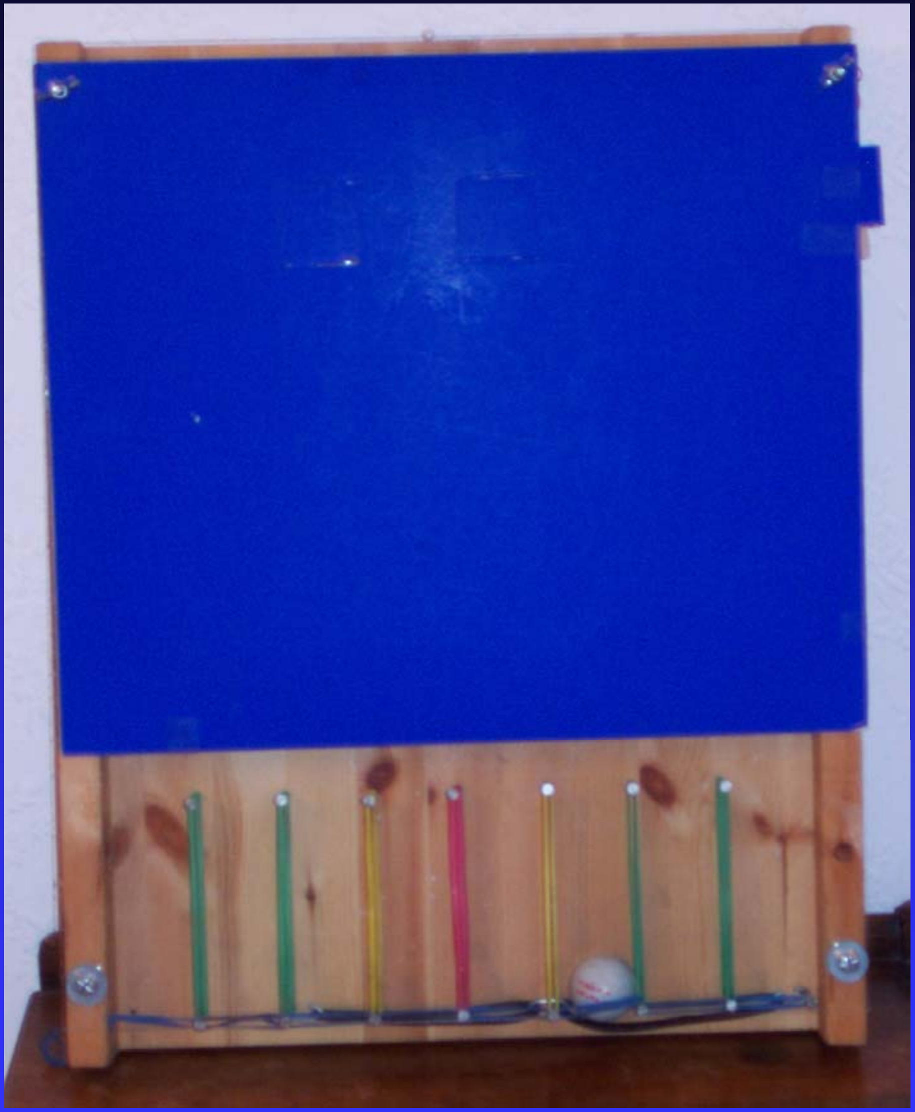
WEATHER CHART, MARCH 31, 1875.



The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.



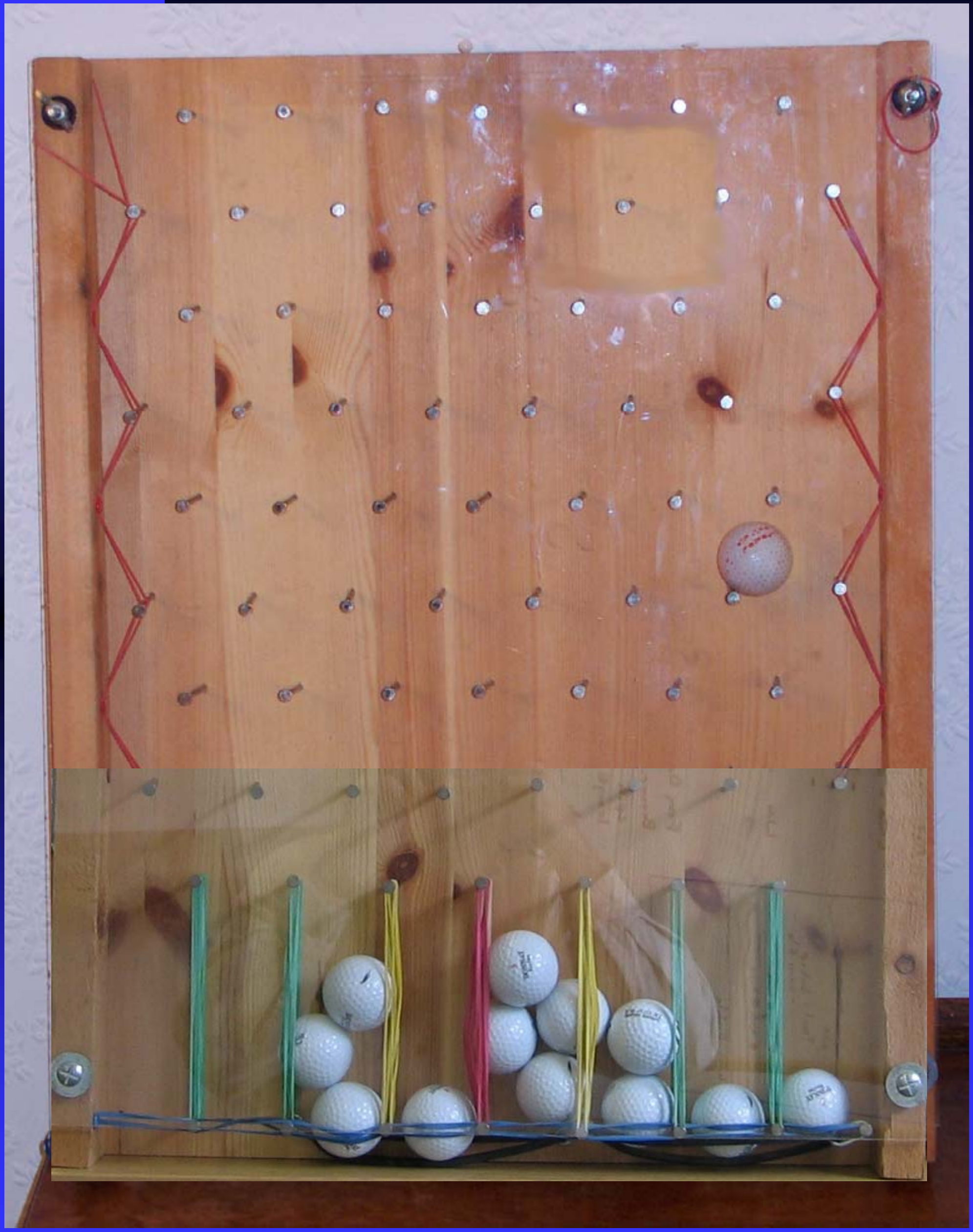
While the Galton Board is a mathematical model…



… this is Not A Galton (NAG) Board.

It is neither stochastic or chaotic; but at least it is!

Lorenz PDF evolution



Reality needn’t be complex,
it merely needs to be real.

What do you see when you look at an ensemble prediction system?

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

Six real-time forecasts based on ensembles…

Now you should each have two pieces of paper with a total three schematic NAG boards on them. This is real data!

I will drop one golf ball at each initial condition

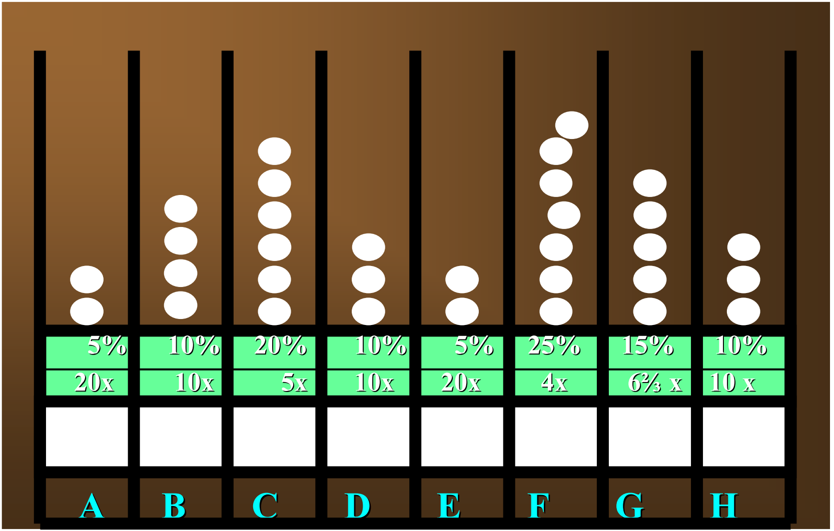
You are asked to

a) Write you name (or some unique identifier) on each page

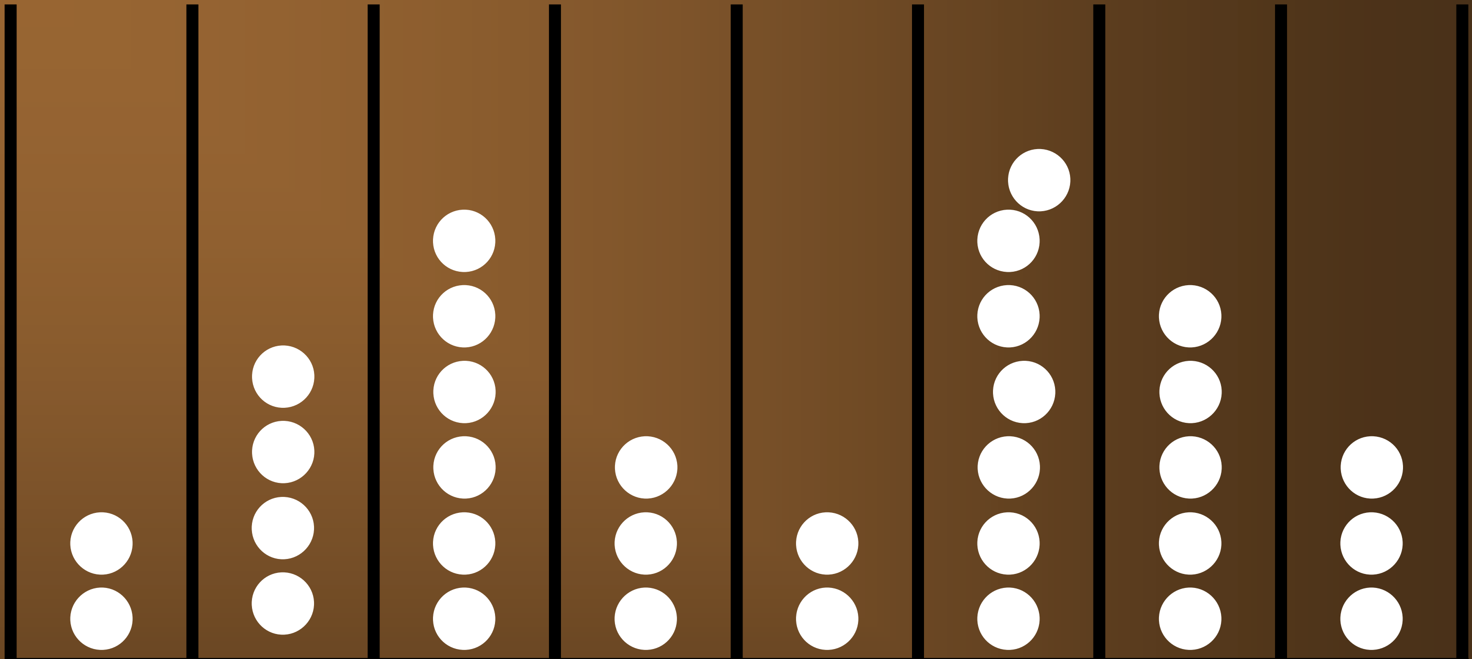
b) Circle the letter you think most likely

c) Distribute a total of 100 e-cents as bets (the odds are given on each bin)

First a dry run…



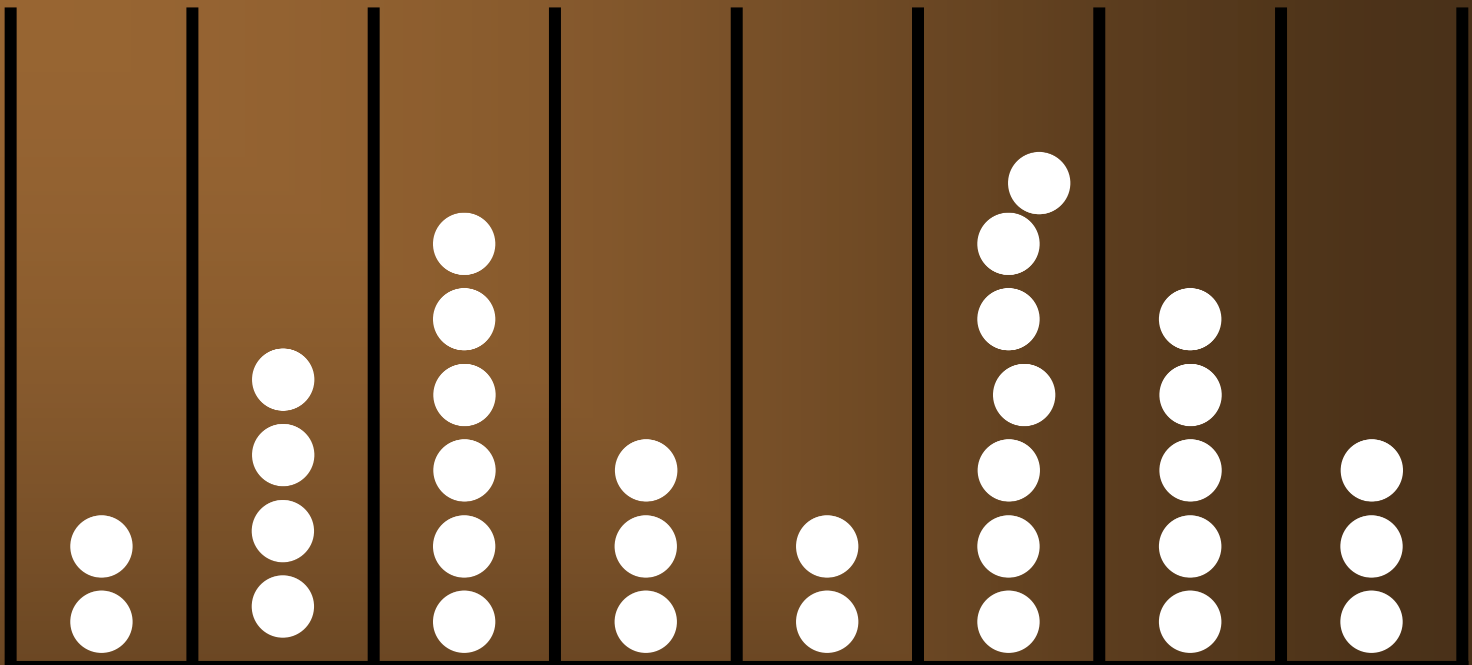
TEST CASE



Prob	5%	10%	20%	10%	5%	25%	15%	10%
Odds	20x	10x	5x	10x	20x	4x	6⅔ x	10 x
Your Bet(s)								
Pick One	A	B	C	D	E	F	G	H

Distribute a total bet of one £ (100 pennies) however you like.
Circle the letter of the bin you think most likely.

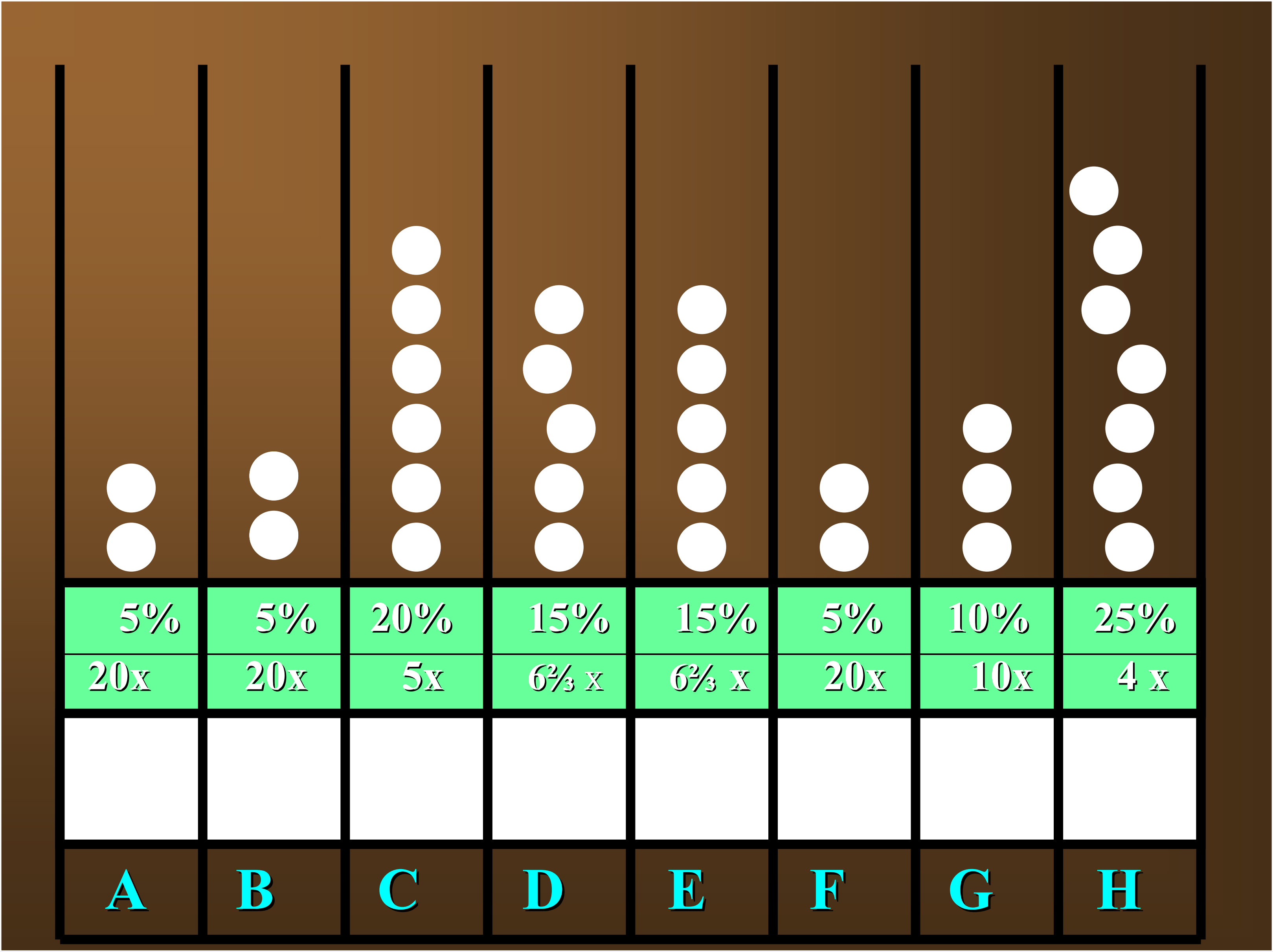
TEST CASE



	A	B	C	D	E	F	G	H
Prob	5%	10%	20%	10%	5%	25%	15%	10%
Odds	20x	10x	5x	10x	20x	4x	6⅔ x	10 x
Your Bet(s)	25	0	0	25	25	0	0	25
Pick One	A	B	C	D	E	[✓ F]	G	H

Distribute a total bet of one £ (100 pennies) however you like.
Circle the letter of the bin you think most likely.

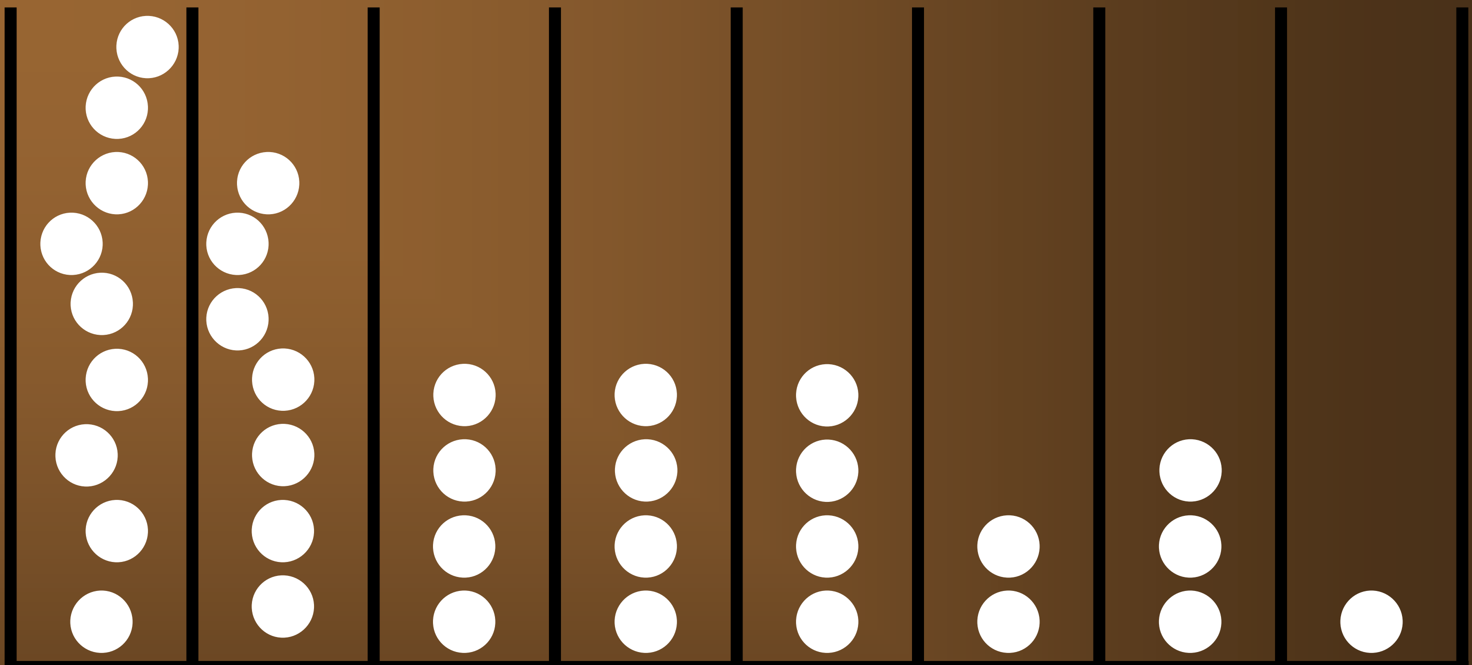
White Pin



Prob	5%	5%	20%	15%	15%	5%	10%	25%
Odds	20x	20x	5x	6⅔ x	6⅔ x	20x	10x	4 x
Your Bet(s)								
Pick One	A	B	C	D	E	F	G	H

Distribute a total bet of one £ (100 pennies) however you like.
Circle the letter of the bin you think most likely.

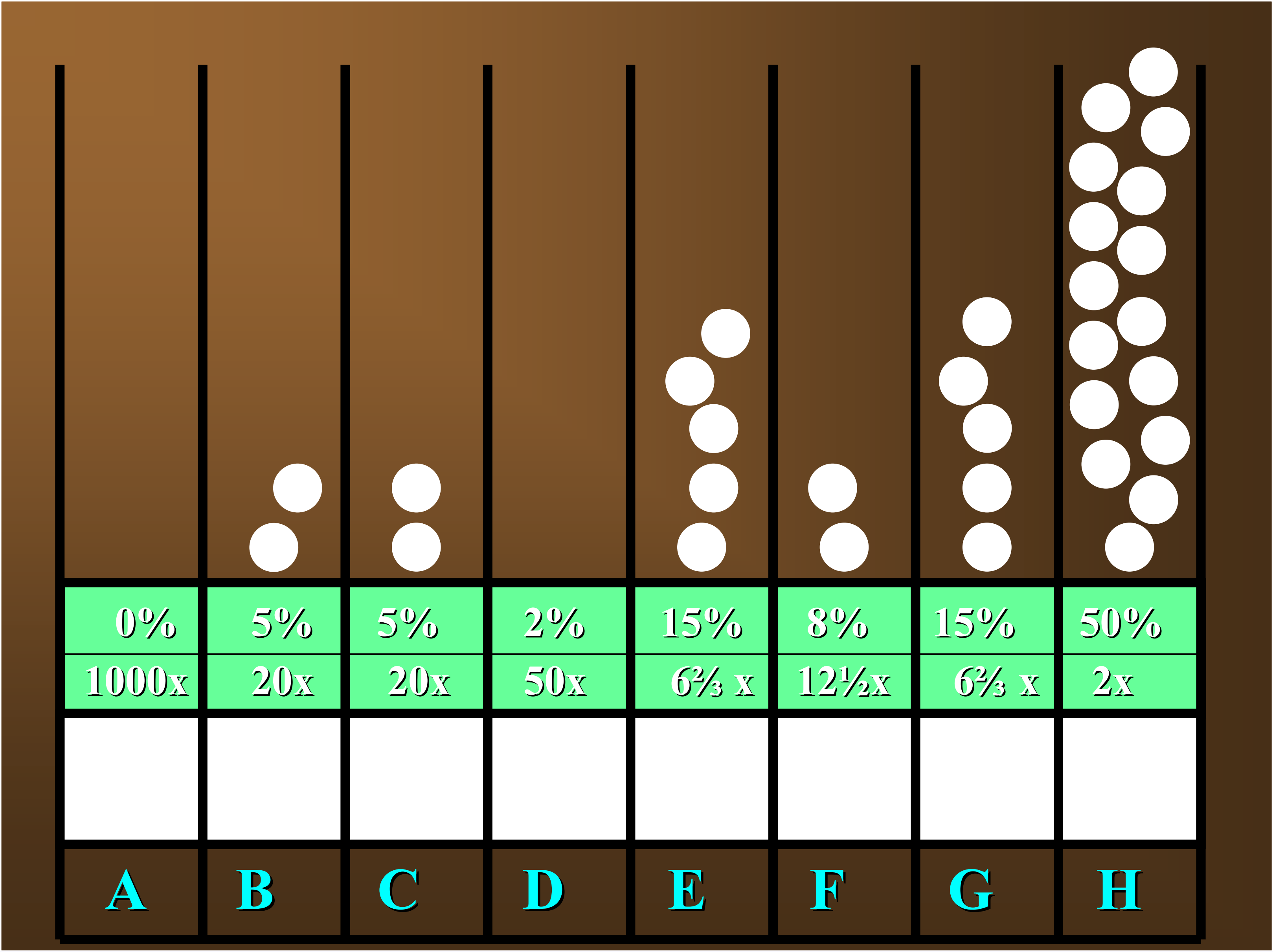
Red Pin
(drop)



Prob	30%	23%	10%	10%	10%	5%	9%	3%
Odds	3⅓x	4⅓x	10x	10x	10x	20 x	11x	33⅓x
Your Bet(s)								
Pick One	A	B	C	D	E	F	G	H

Distribute a total bet of one £ (100 pennies) however you like.
Circle the letter of the bin you think most likely.

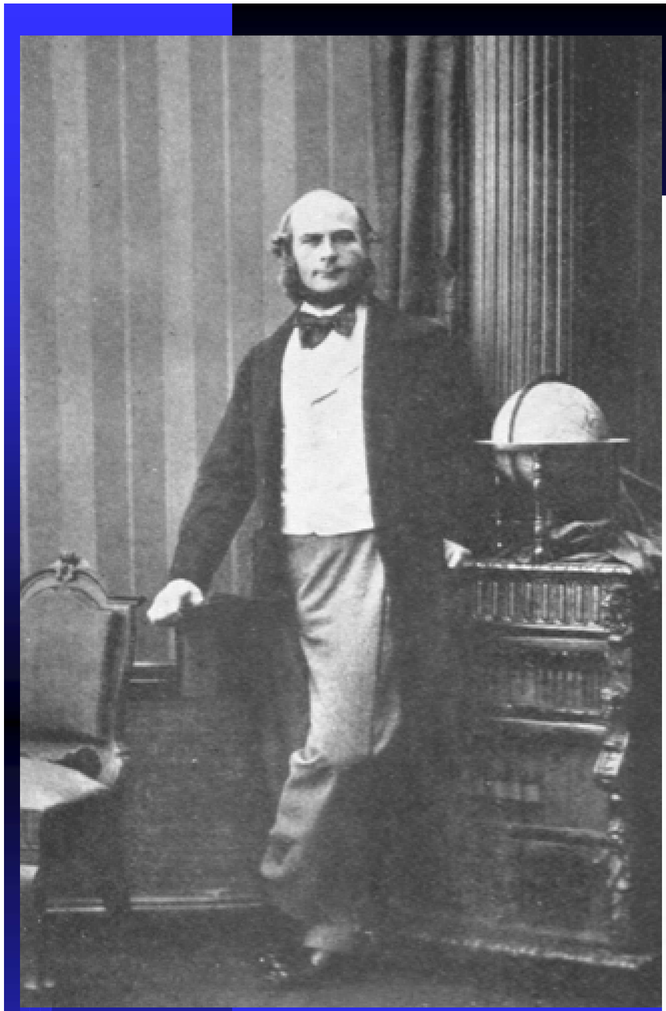
Yellow Pin



	A	B	C	D	E	F	G	H
Prob	0%	5%	5%	2%	15%	8%	15%	50%
Odds	1000x	20x	20x	50x	6⅔ x	12½x	6⅔ x	2x
Your Bet(s)								
Pick One	A	B	C	D	E	F	G	H

Distribute a total bet of one £ (100 pennies) however you like.
Circle the letter of the bin you think most likely.

This Galton Board is a mathematical model.



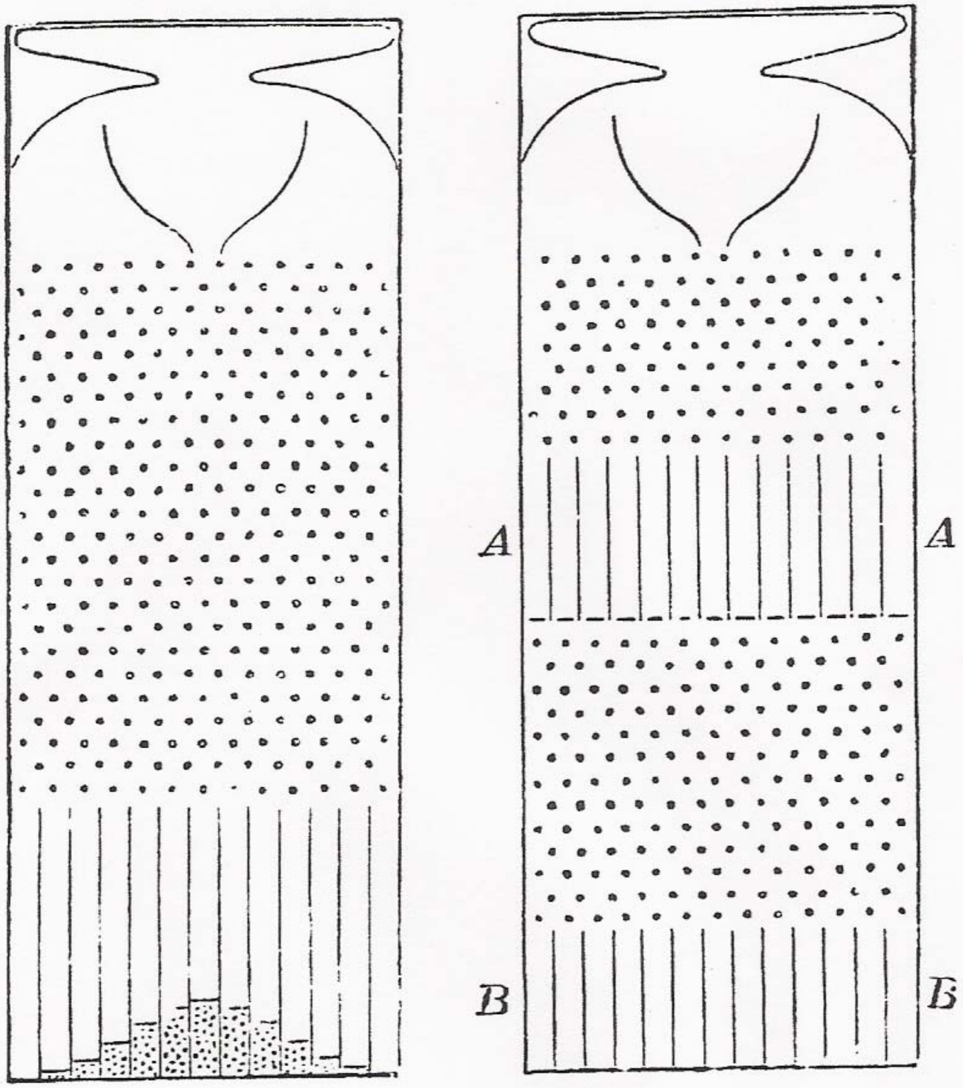
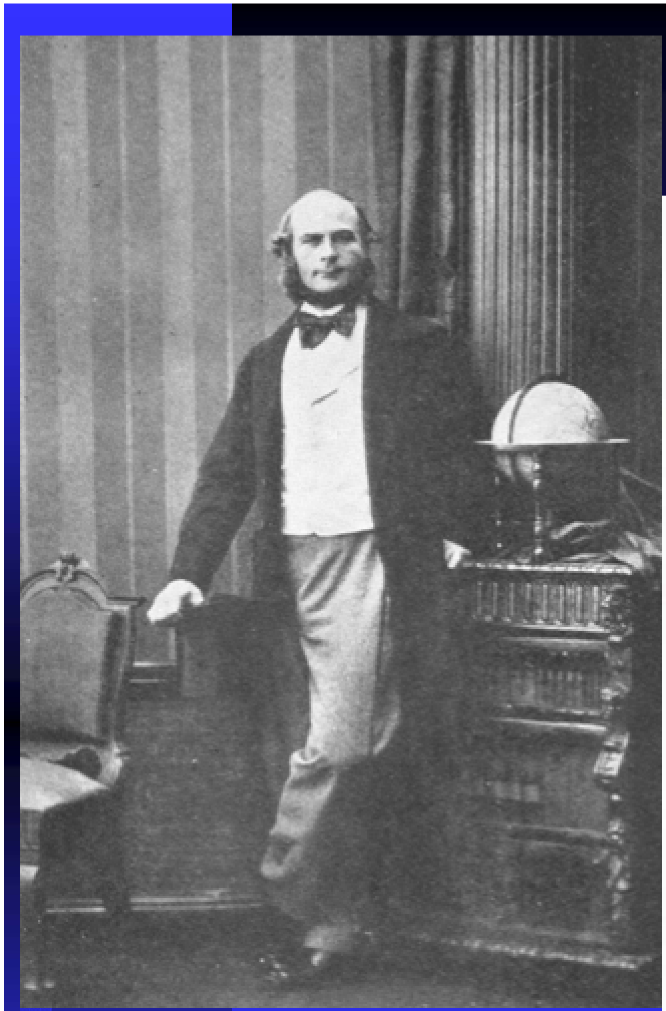
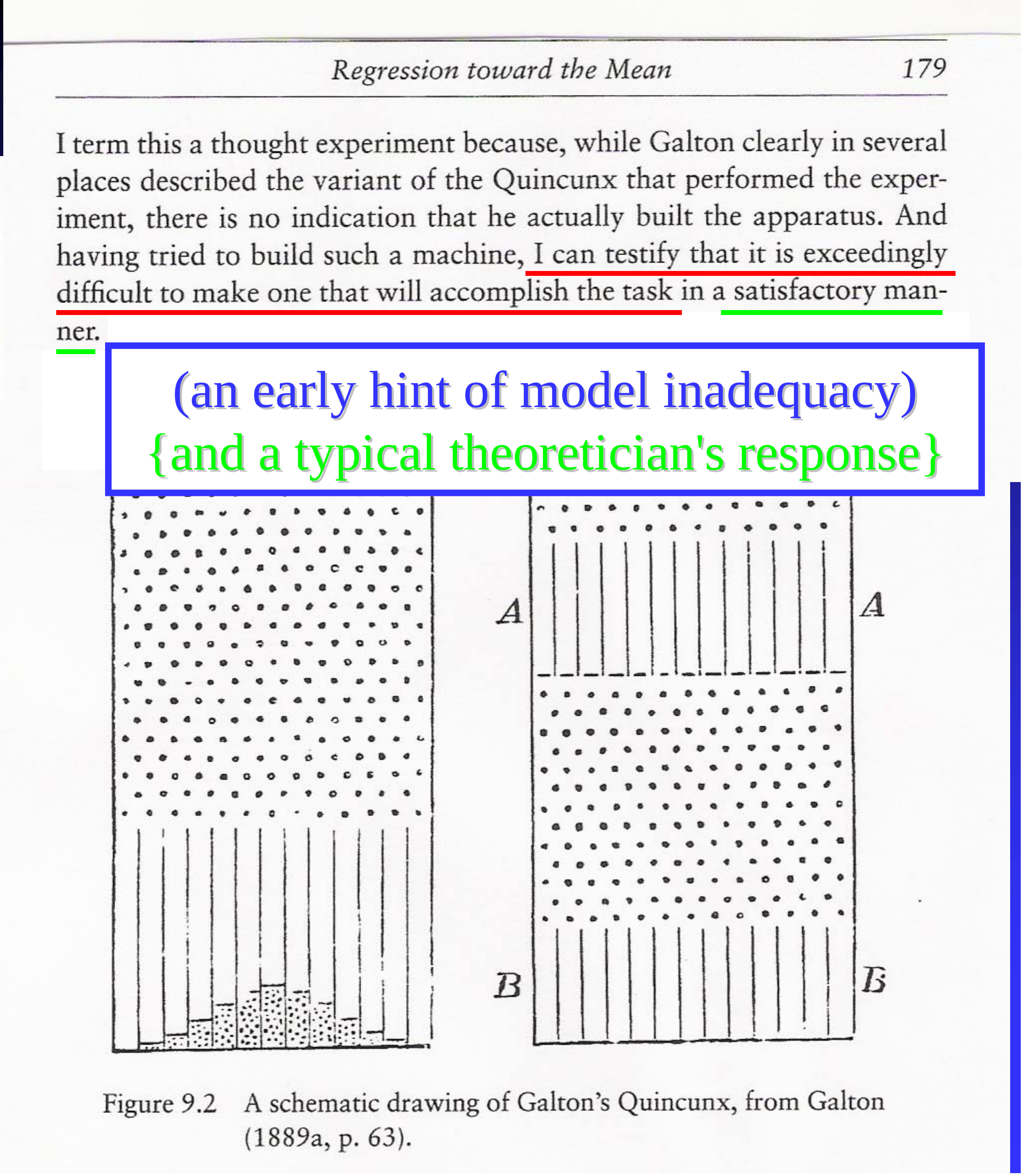


Figure 9.2 A schematic drawing of Galton's Quincunx, from Galton (1889a, p. 63).





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.

(an early hint of model inadequacy)

{and a typical theoretician's response}

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

In the first three forecasts the model was perfect, and the ensemble was perfect.

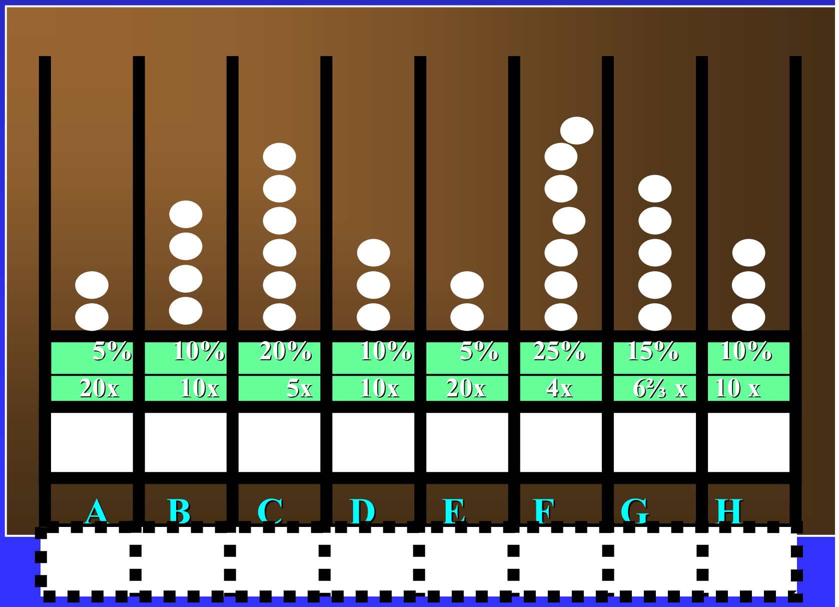
Go back to the schematic NAG board marked “White Pin”
This time we forecast reality, and reality is NOT a golf ball

The red rubber ball of Reality will fall once from each initial condition

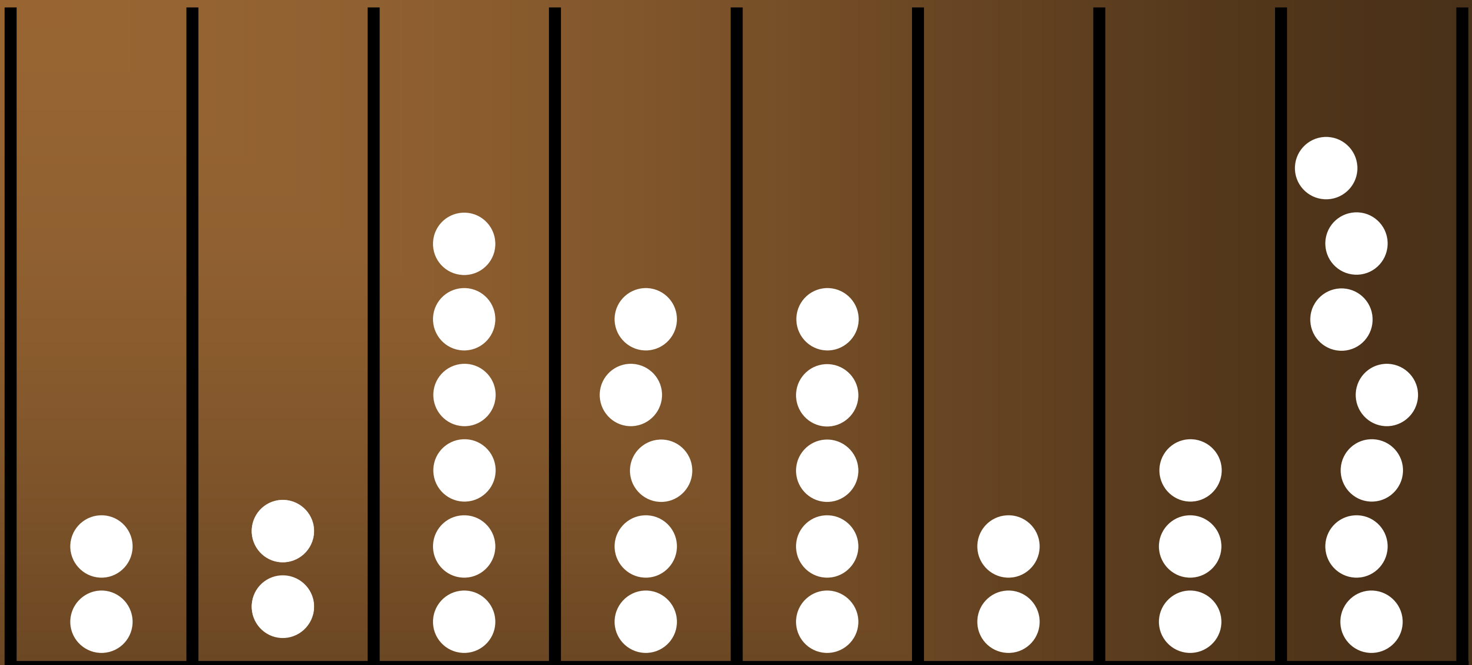
Before it does, you are asked to use the dotted boxes to

a) Distribute a total of 100 e-cents as bets (the odds are given on each bin)
b) Place an X in the bin you think most likely

There is no dry run with reality…

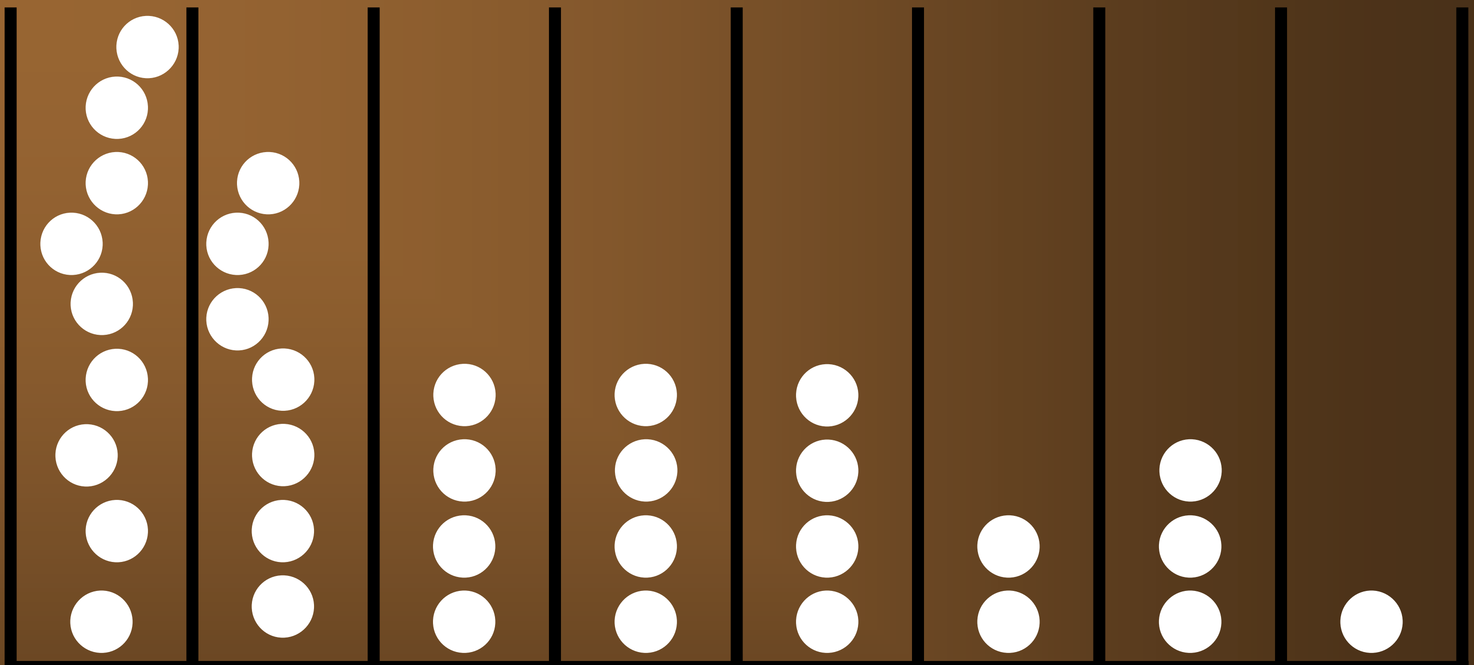


White Pin



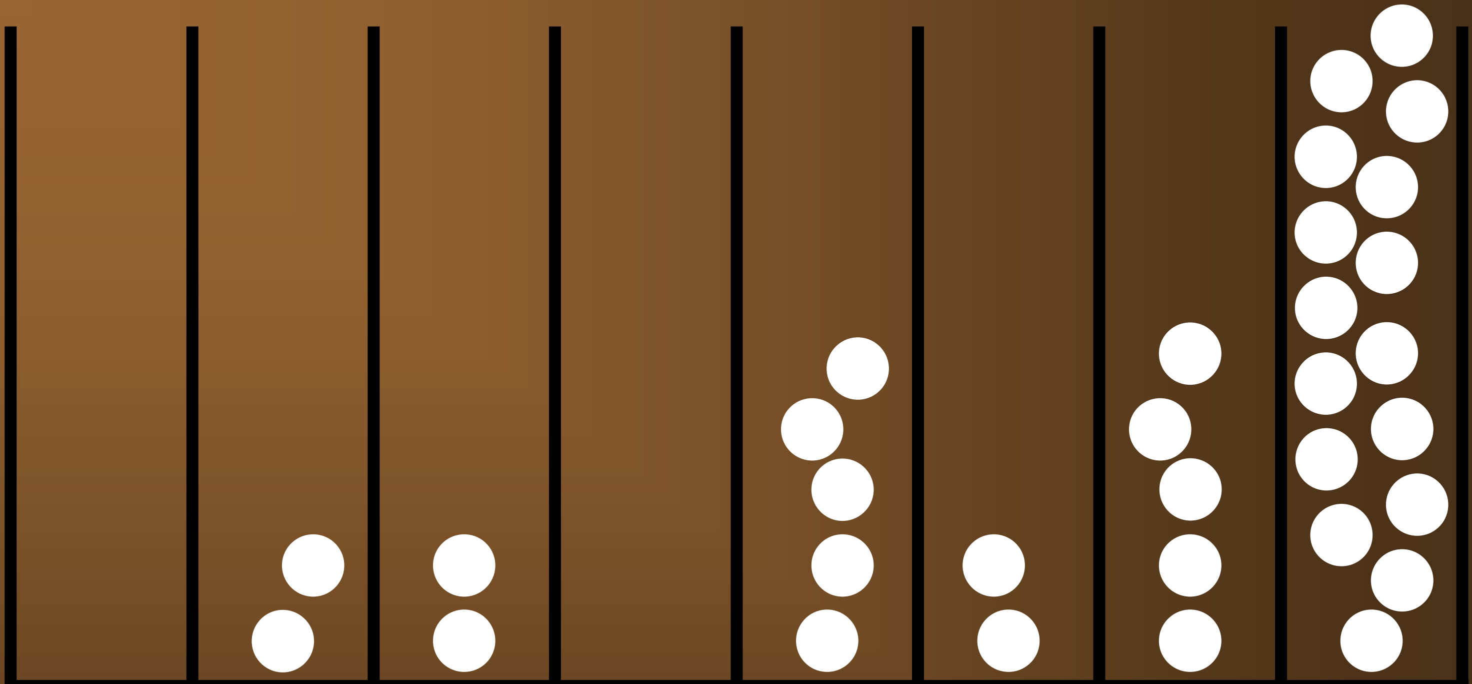
	A	B	C	D	E	F	G	H
Prob	5%	5%	20%	15%	15%	5%	10%	25%
Odds	20x	20x	5x	6⅔ x	6⅔ x	20x	10x	4 x
Your Bet(s)								
Pick One								

Red Pin
(drop)



Prob	30%	23%	10%	10%	10%	5%	9%	3%
Odds	3⅓x	4⅓x	10x	10x	10x	20x	11x	33⅓x
Your Bet(s)								
Pick One	A	B	C	D	E	F	G	H

Yellow Pin



	A	B	C	D	E	F	G	H
Prob	0%	5%	5%	2%	15%	8%	15%	50%
Odds	1000x	20x	20x	50x	6⅔ x	12½x	6⅔ x	2x
Your Bet(s)								
Pick One								

Who changed their bets on the second round (Reality)? Why?

What sort of verification tools would help you with the Red Rubber Ball, noting that we never see the same initial condition twice?

We'll not see two similar medium range PDFs before the sun dies.

Many industrial users of weather forecasts play this game using the ECMWF ensemble prediction system every day, with much higher stakes.
What verification measures do they want? Need?

Coffee

Please hand in the sheets to Du

Coffee!

(and questions)

KEN FISHER ON WHY TOO FEW AUDITORS IS BAD FOR THE STOCKMARKET

Bloomberg Money

[redacted]

Issue No. 96
£3.85

THIS MONTH	
FUND PICKERS	2
DIVIDENDS	44
PROPERTY	5
PENSIONS	57
ANNUITIES	5
SAVINGS	6
AGENDA	64-6
STATISTICS	70-8

WIN A LUXURY HOTEL BREAK FOR TWO

A WEALTH OF INFORMATION FOR THE PRIVATE INVESTOR

RETIREMENT INCOME: WHY A SIPP WILL BE YOUR FLEXIBLE FRIEND

TRADING PLACES: SEVEN STEPS TO SUCCESSFUL SHARE DEALING

EURO-VISION: WE TALK TO FIDELITY'S MAN IN EUROPE

HIT PARADE: THE WINNERS OF OUR ANNUAL AWARDS REVEALED

AWARD WINNING PREDICTIONS

Our fund managers of the year look ahead to the next 12 months

2007

PLUS ALL THE LATEST FROM OUR REGULAR AWARD WINNING COLUMNISTS

When should a prediction win an award?

How many fund managers do you need for one of them to make a big profit?

-or-

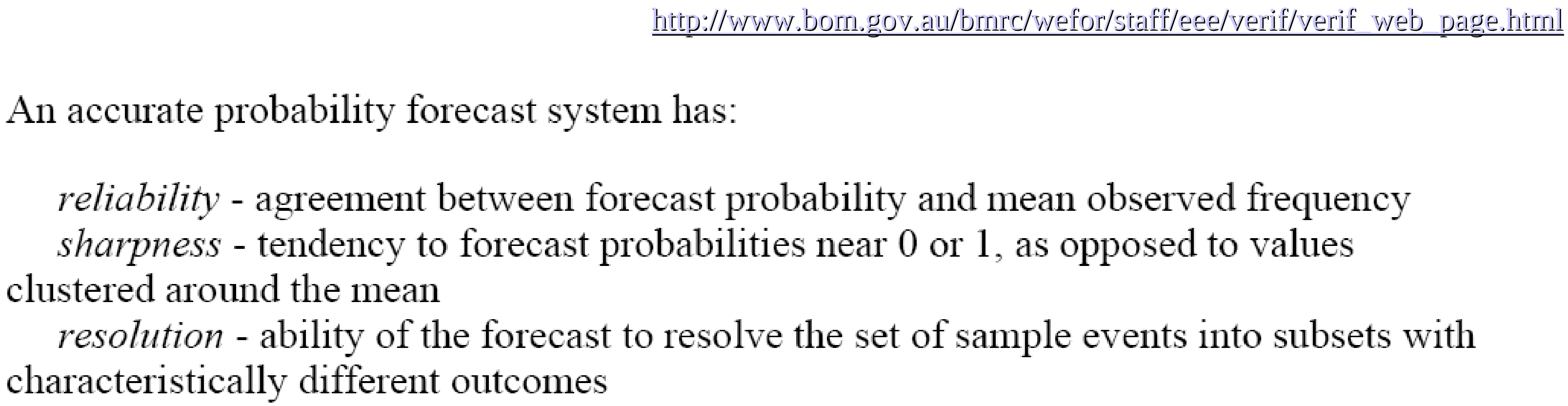
Whenever there are lots of forecasters forecasting, identifying insight gets harder whatever the verification measures.

“Past performance is no guarantee of future returns.”

Use common sense too!

The NAG Board Forecasts…

Uncertainty, Utility, Adequacy, and Verification



http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

An accurate probability forecast system has:

reliability - agreement between forecast probability and mean observed frequency
sharpness - tendency to forecast probabilities near 0 or 1, as opposed to values clustered around the mean
resolution - ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes

In short: Is this bar too high?

Accurate probabilities would be great to have, but are they a requirement for a useful forecast system?
(and might reaching for them make our ensembles less useful?)
Is there a principled approach to EPS improvement?

Yesterday you had a lecture on

"Confidence Intervals and Hypothesis tests".

I think there are some deep similarities between well known difficulties in hypothesis testing and as yet unsolved difficulties in verification.

> Results from big samples are nearly always significant even when the effects are quite small in magnitude
>
> Null hypotheses are often rather silly and obviously untrue.

C Chatfield, Problem Solving: A Statistician's Guide. pg *79*

What does this mean for verification?

Results from big samples are nearly always significant even when the effects are quite small in magnitude

Null hypotheses are often rather silly and obviously untrue.

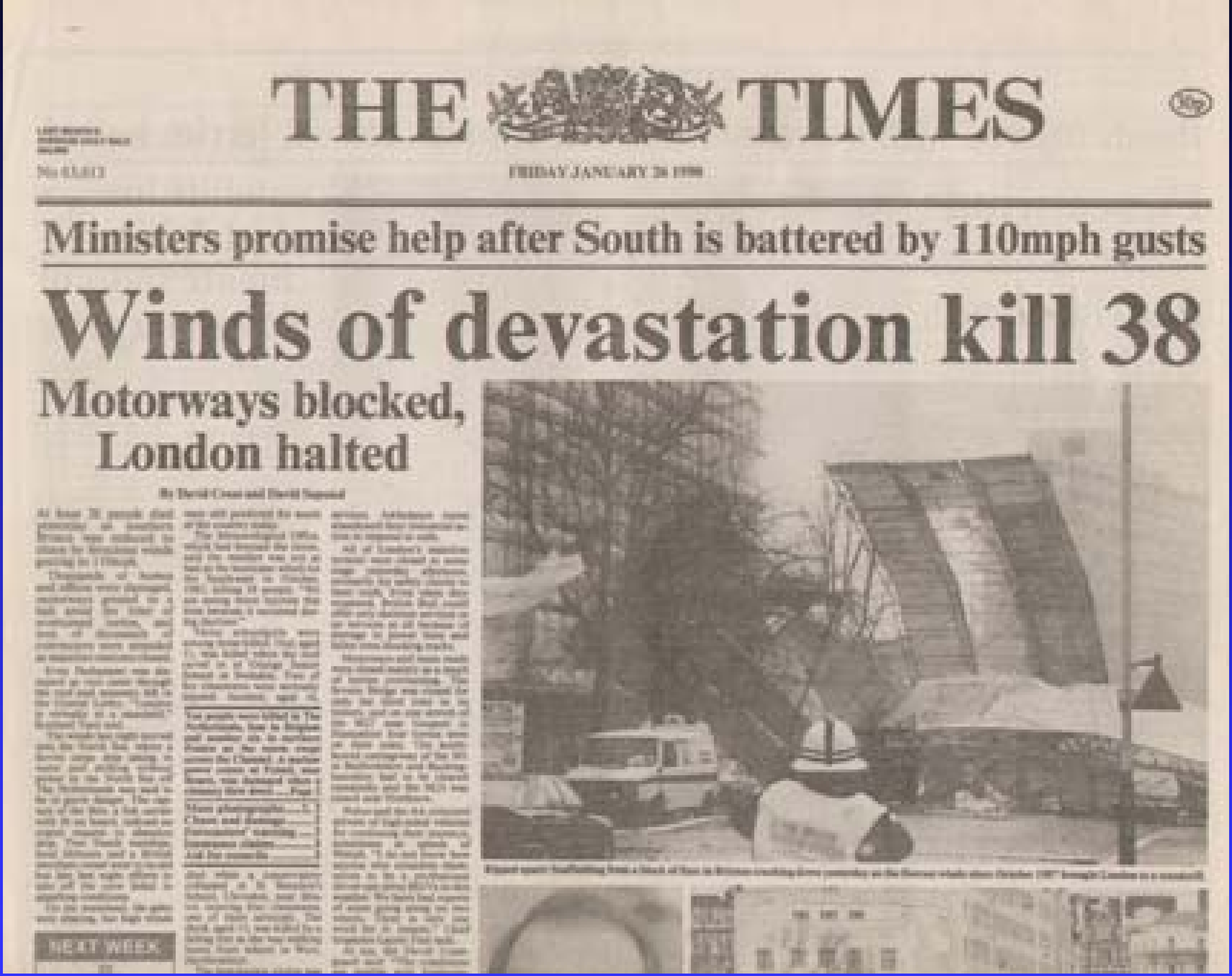
C Chatfield, Problem Solving: A Statistician's Guide. ***pg 79***

We often test the null that our probability forecasts do indeed reflect the probability of the target.

This is "silly and obviously untrue": our models are imperfect and at present we do not even attempt to form ensembles that would give an accountable probability *even if* our model was perfect!

Can verifying useful models as probability forecasts make them less useful? (just as tuning a perfect model with RMS would?)

Does "distance from" being an accurate probability forecast reflect utility?



THE TIMES

FRIDAY JANUARY 26 1990

Ministers promise help after South is battered by 110mph gusts

Winds of devastation kill 38

Motorways blocked, London halted

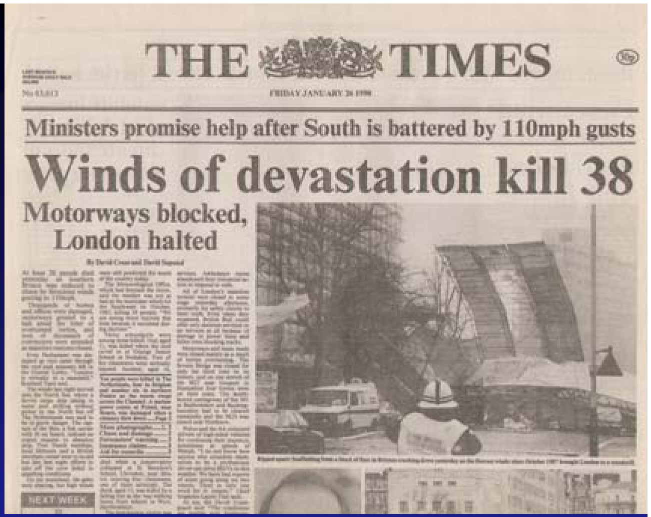
By David Cross and David Sapsted

[illegible]

NEXT WEEK

Burns Day Storm, 1990

Burns Day Storm, 1990



THE TIMES

FRIDAY JANUARY 26 1990

Ministers promise help after South is battered by 110mph gusts

Winds of devastation kill 38

Motorways blocked, London halted

1990: This was pre-ensembles

Critical Observations from two ships

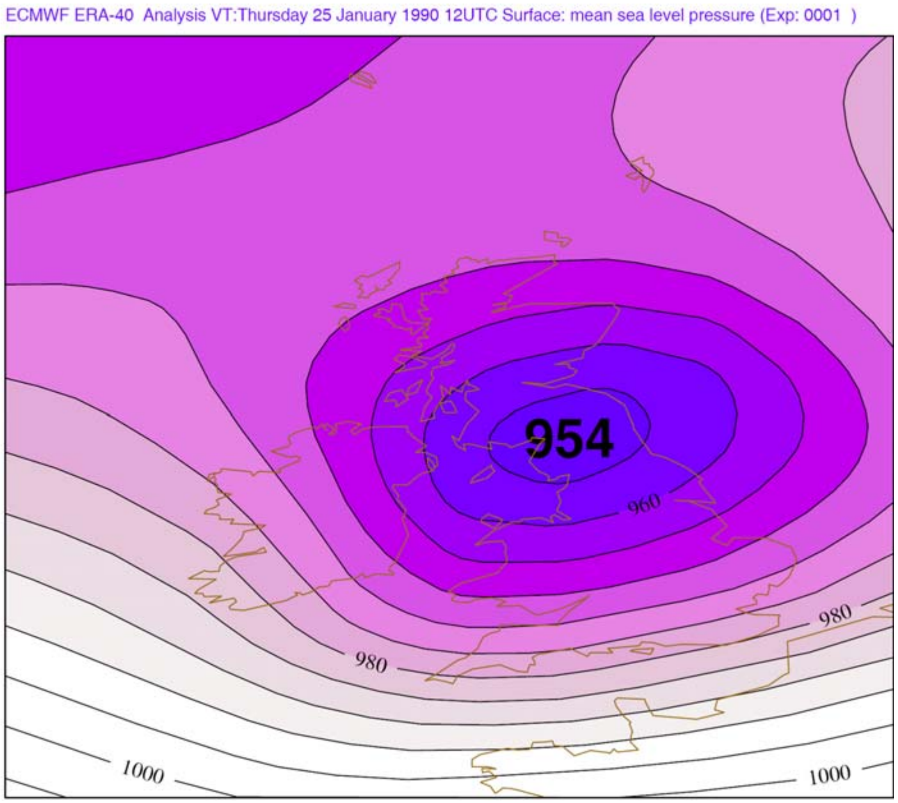
Rejected by UKMO QC system

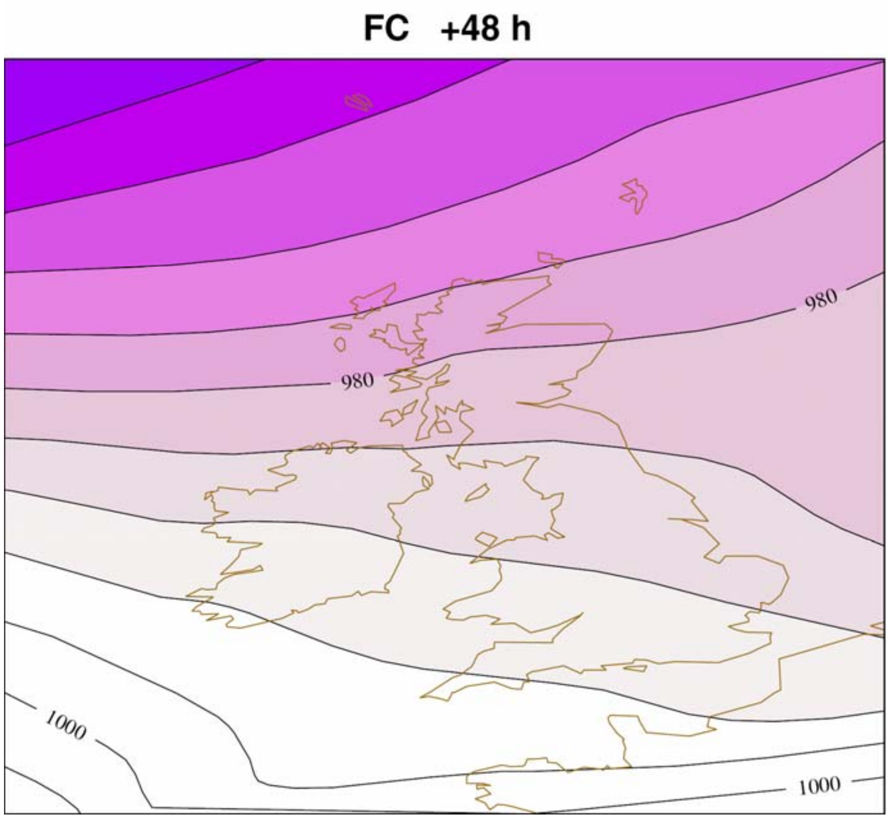
Reinserted by “Intervention Forecaster”

who knew s\he would only get to see one model run!

Well-forecast storm at 24 hours lead time.

Significant socio-economic value in this single forecast!


ECMWF ERA-40 Analysis VT:Thursday 25 January 1990 12UTC Surface: mean sea level pressure (Exp: 0001)
954
960
980
980
1000
1000



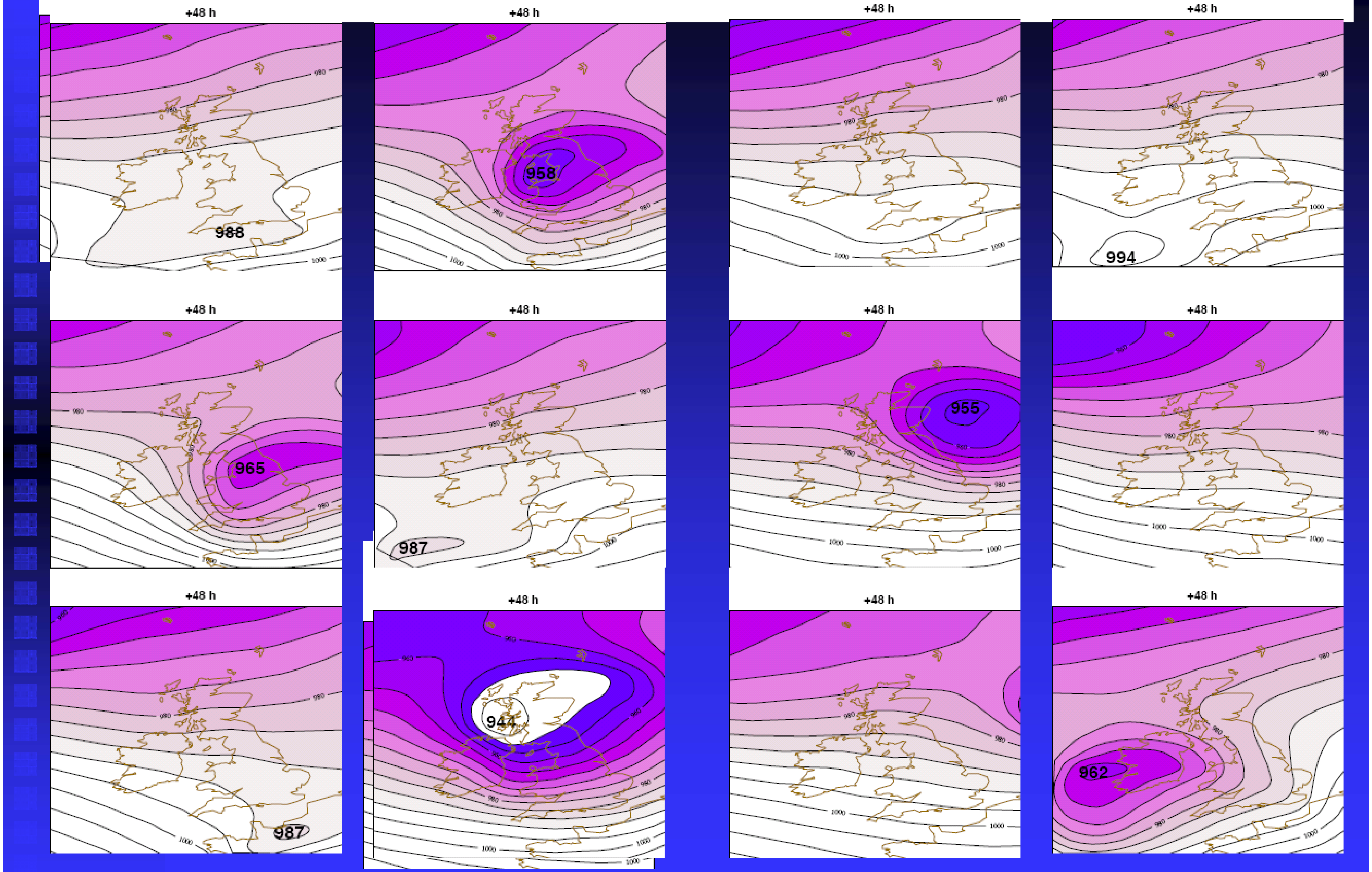
Figures from Martin Leutbecher, ECMWF

(Re-) Analysis

"Best First Guess" at 48 hours
BFG Forecast using a "2002" model.

But in 2002 we had an ensemble…

ECMWF 48 Hour lead time EPS for Burns Day: high utility without a PDF.



What information is in this distribution of golf balls; and how are we to use it?

Thanks to Martin Leutbecher, ECMWF

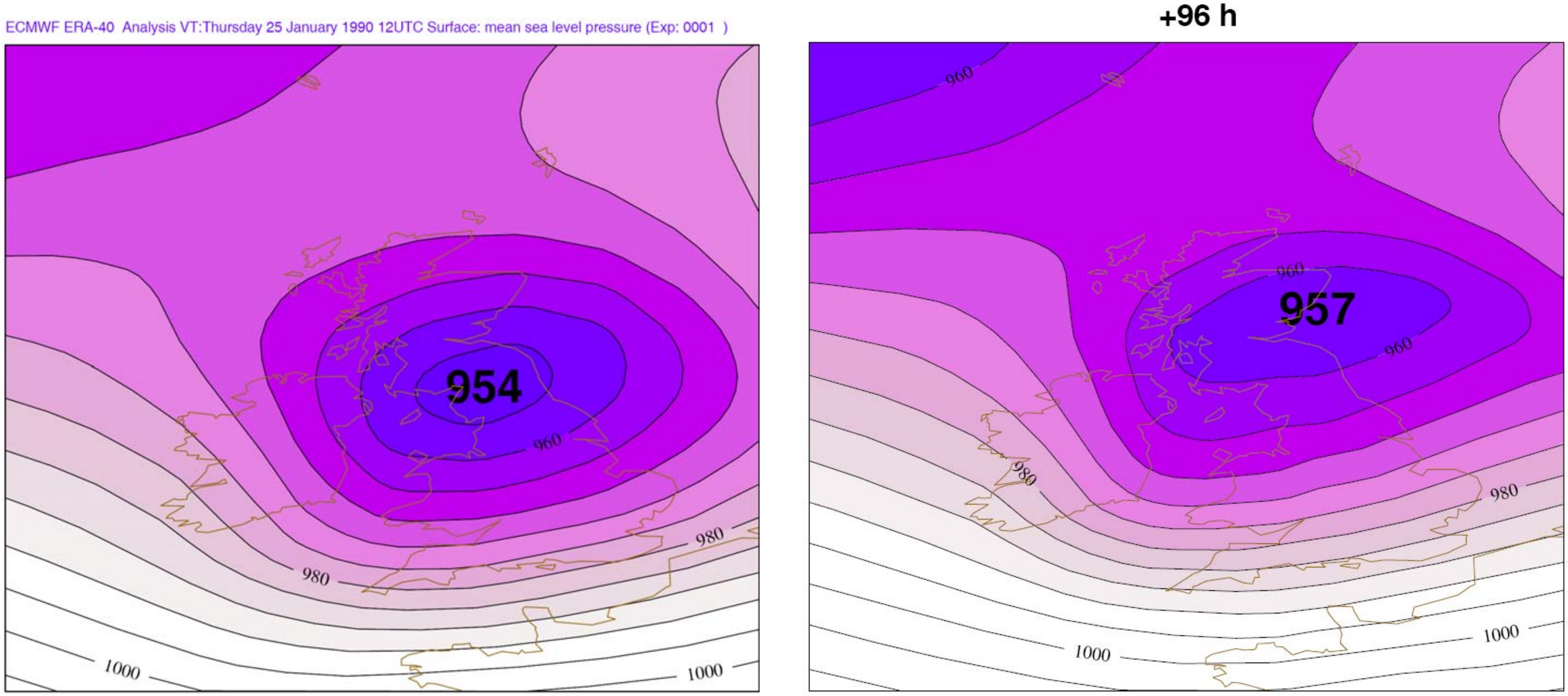
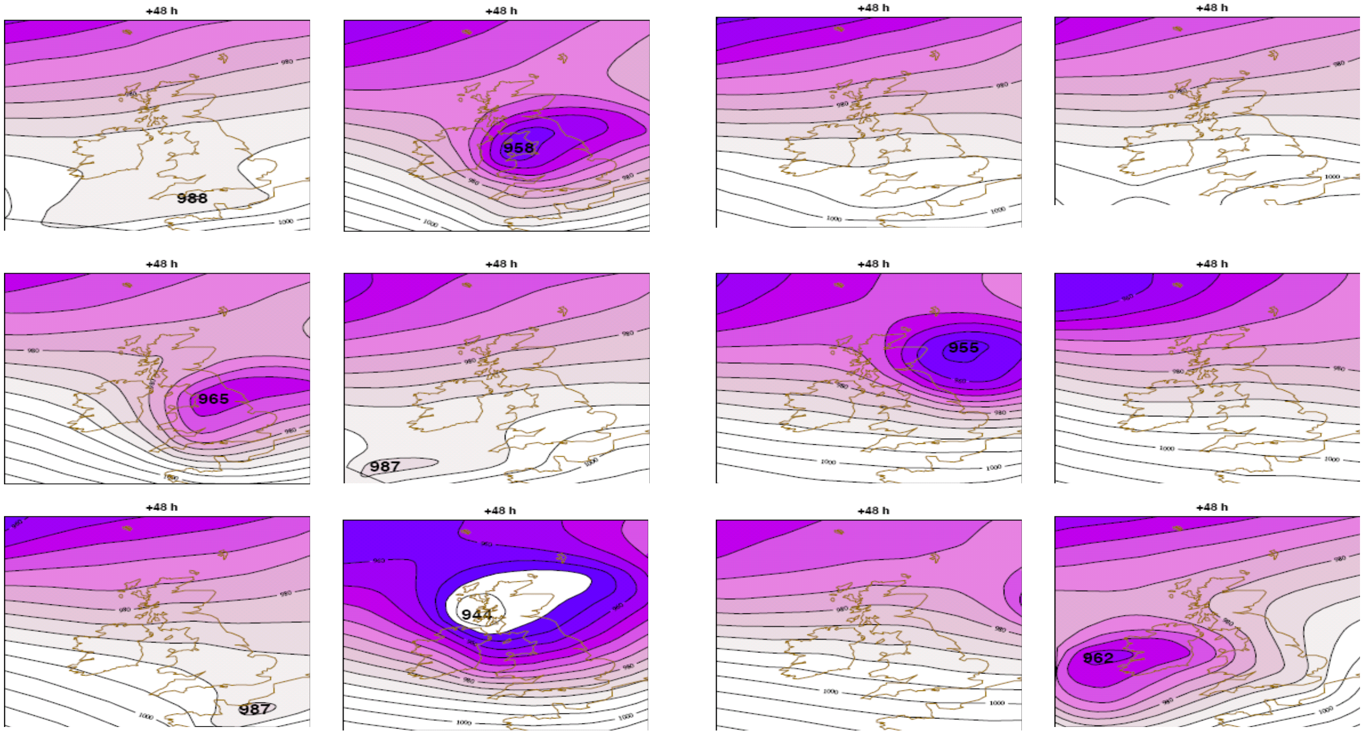

ECMWF ERA-40 Analysis VT:Thursday 25 January 1990 12UTC Surface: mean sea level pressure (Exp: 0001)
954
960
980
980
1000
1000
+96 h
957
960
960
960
980
980
1000
1000

Figure from Martin Leutbecher, ECMWF

Just to stress the point: some ensembles gave good early warning four days ahead.

(of course, we need to make sure the EPS does not make storms over Scotland every day of the winter…)

ECMWF 48 Hour lead time EPS for Burns Day: high utility without a PDF.



Thanks to Martin Leutbecher and ECMWF

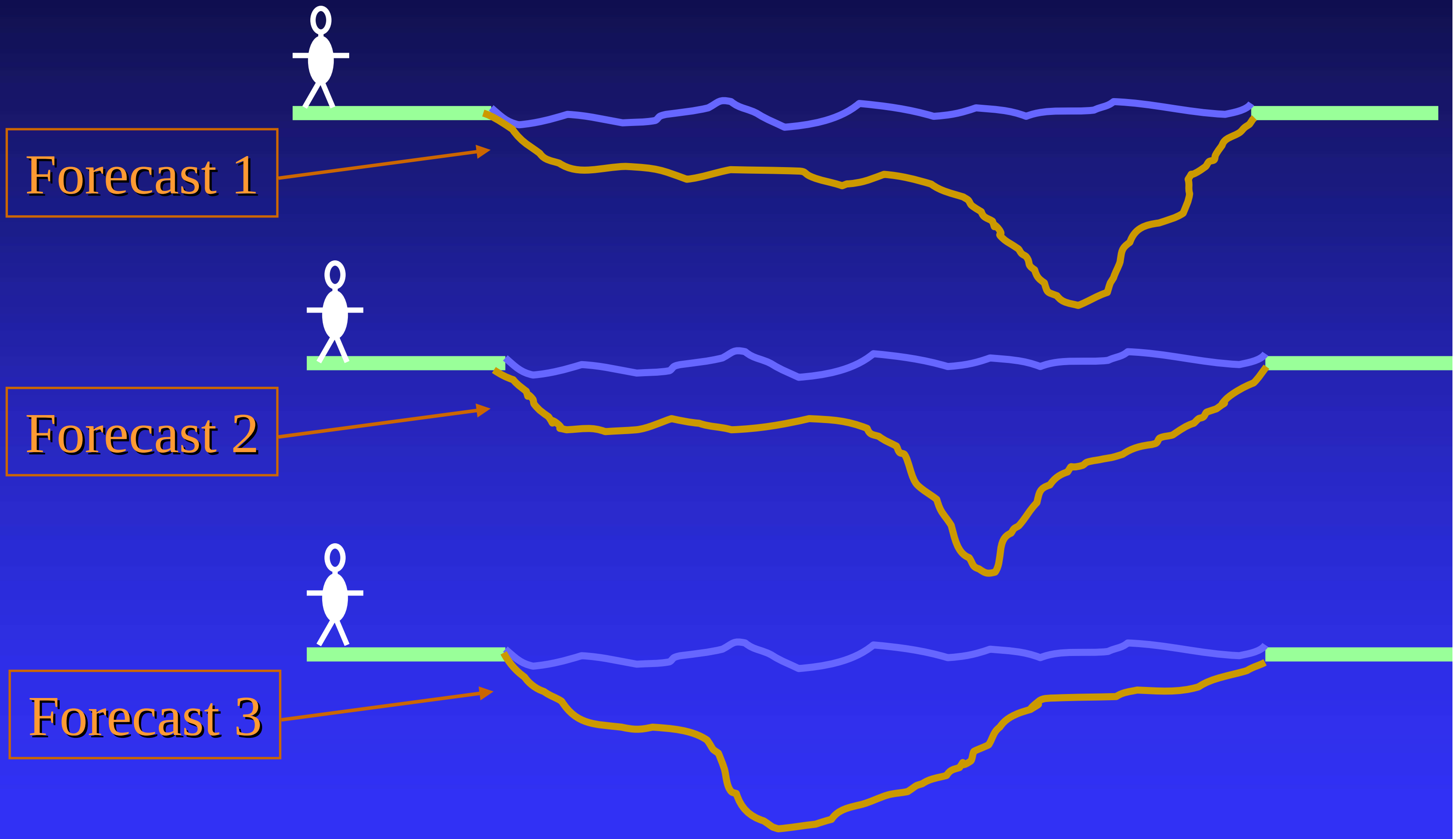
ECMWF ensembles contain valuable information, we must be careful not to destroy or discard it!

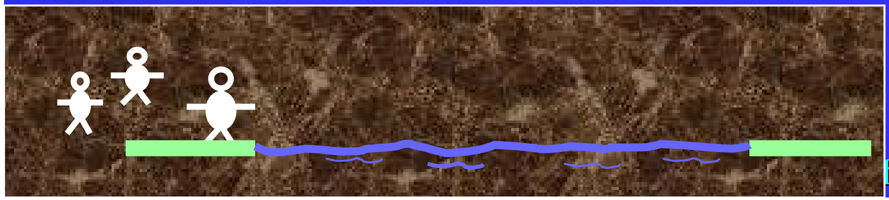
But how might we verify it?

The parable of the three statisticians.

Three non-Floridian statisticians come to a river, they want to know if they can cross safely. (They cannot swim.)

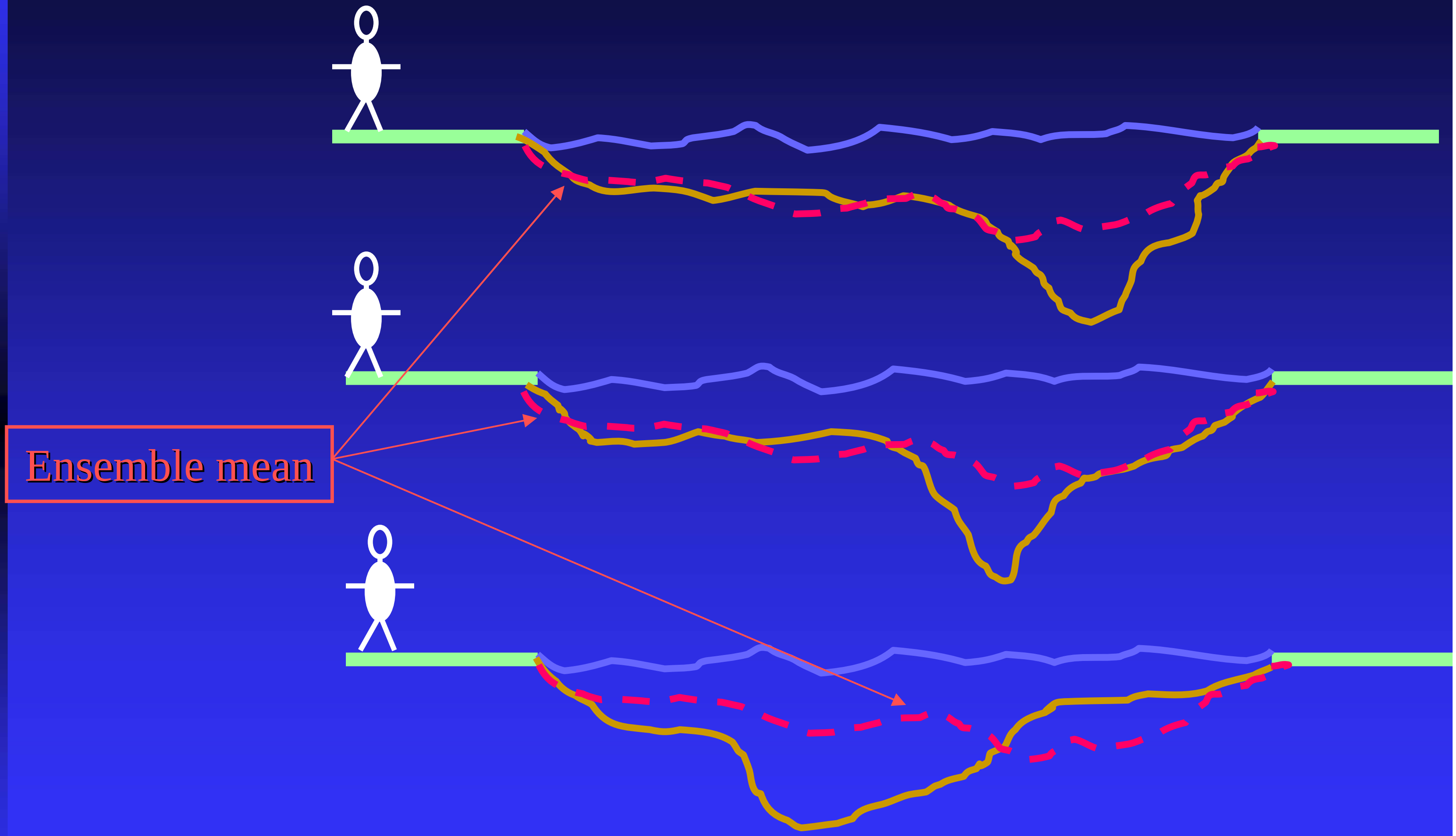
Three non-Floridian statisticians wish to cross a river.
Each has a forecast of depth which indicates they will drown.



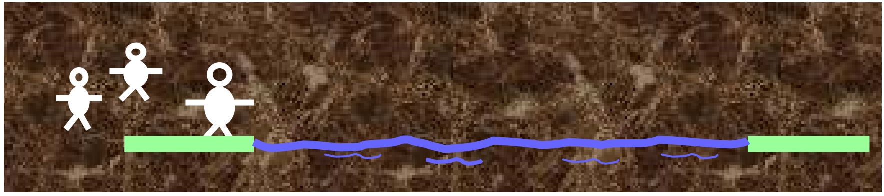


So they have an ensemble
forecast,with three members

Three non-Floridian statisticians wish to cross a river.
Each has a forecast of depth which indicates they will drown.
So they average their forecasts and decide based on the ensemble mean…



Ensemble mean



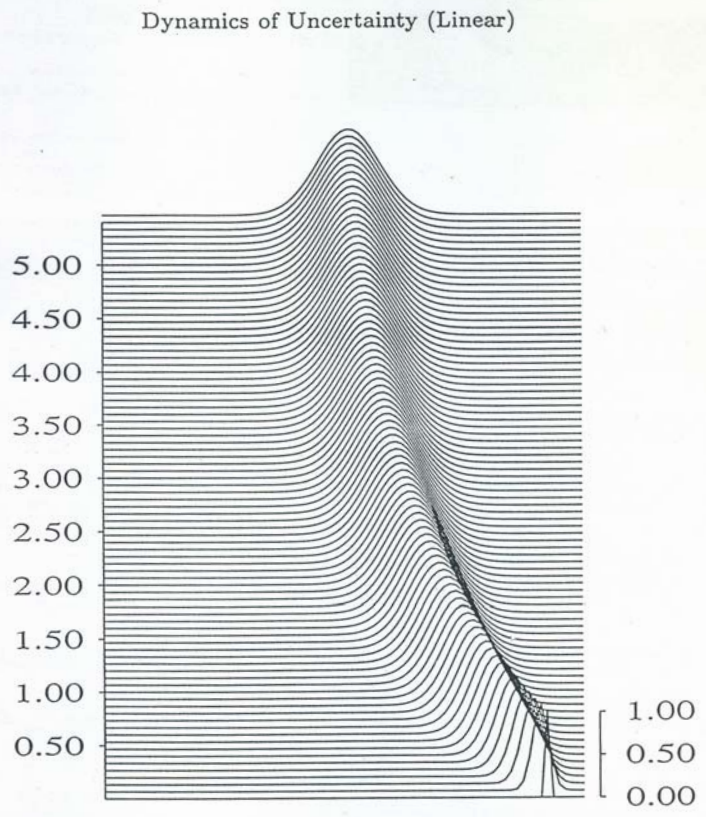
Is this a good idea?

No!

Ensembles contain information, we must be careful not to destroy or discard it!

How do we even address these questions?

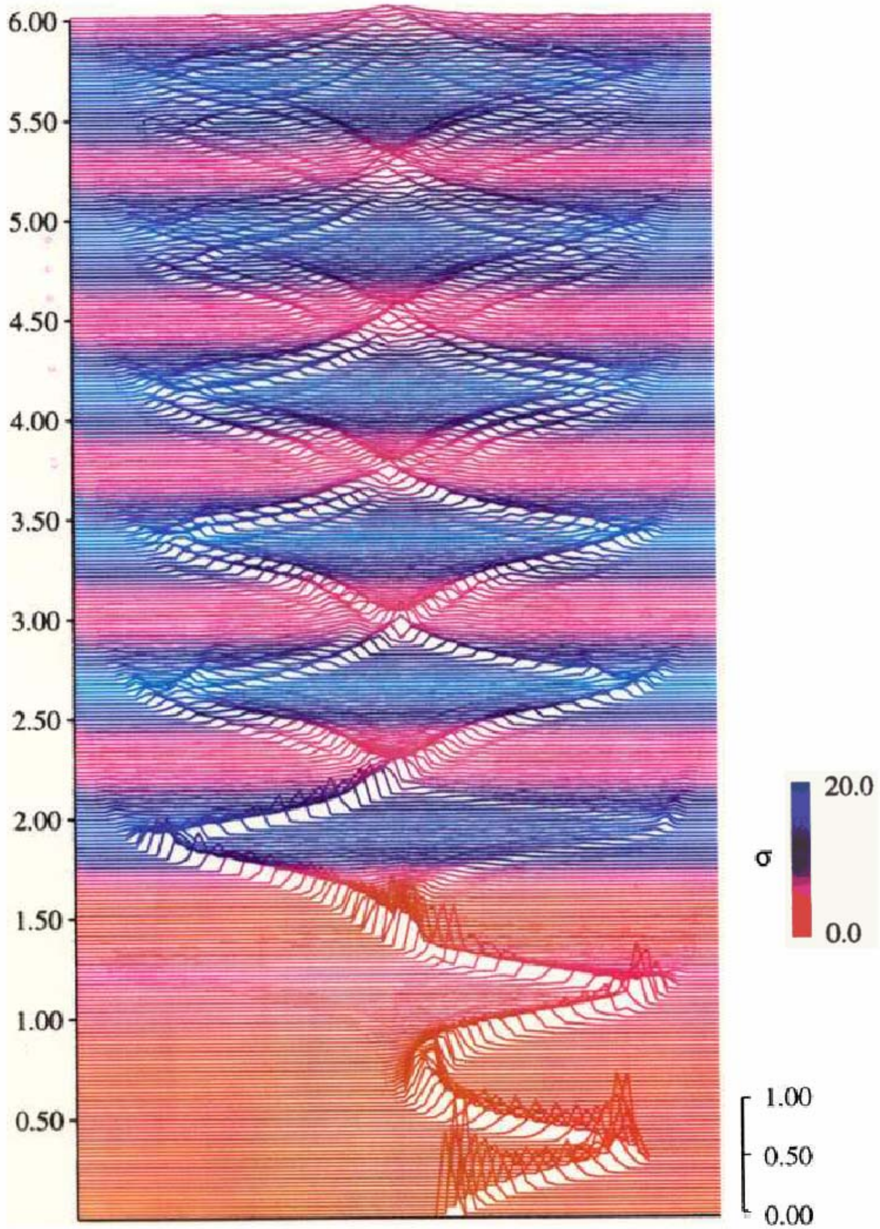
Probability Forecasts



AR(1) Yule (1926)

In linear systems almost everything is normally distributed, knowing the mean and the variance tell you everything *in that case*.

Many verification tools were developed for this case.



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

This is a simple nonlinear case.

So does any of this really matter? Why can't we just use the RMS distance from the ensemble mean and carry on…

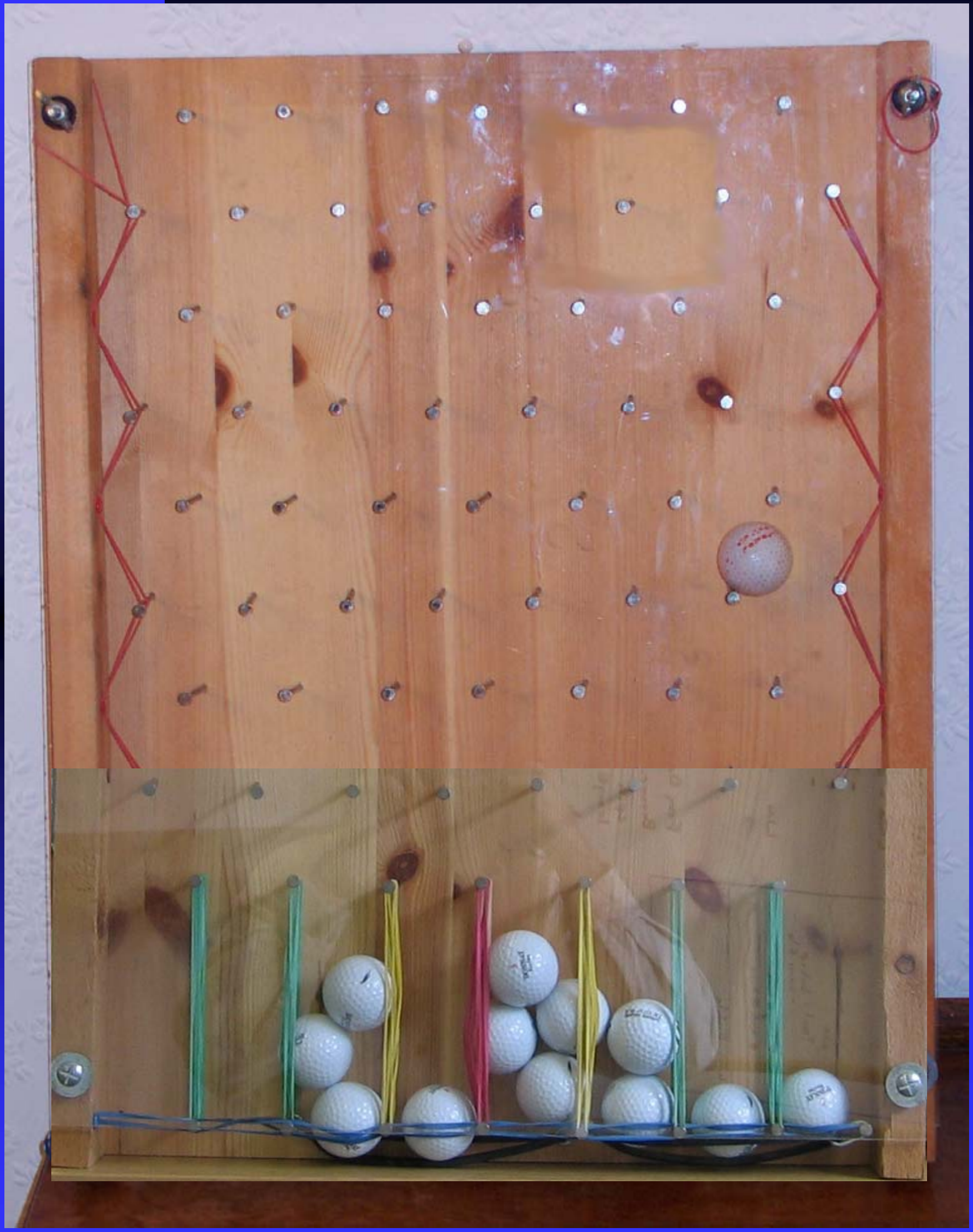
(McSharry & Smith, PRL 1999)

Tuning parameters based upon RMS verification will reject the parameters that generated the data in a perfect model!

IGN simultaneously evaluates the model and the ensemble.

And in this case we have a perfect model; what do we do if our models disagree?

Lorenz PDF evolution

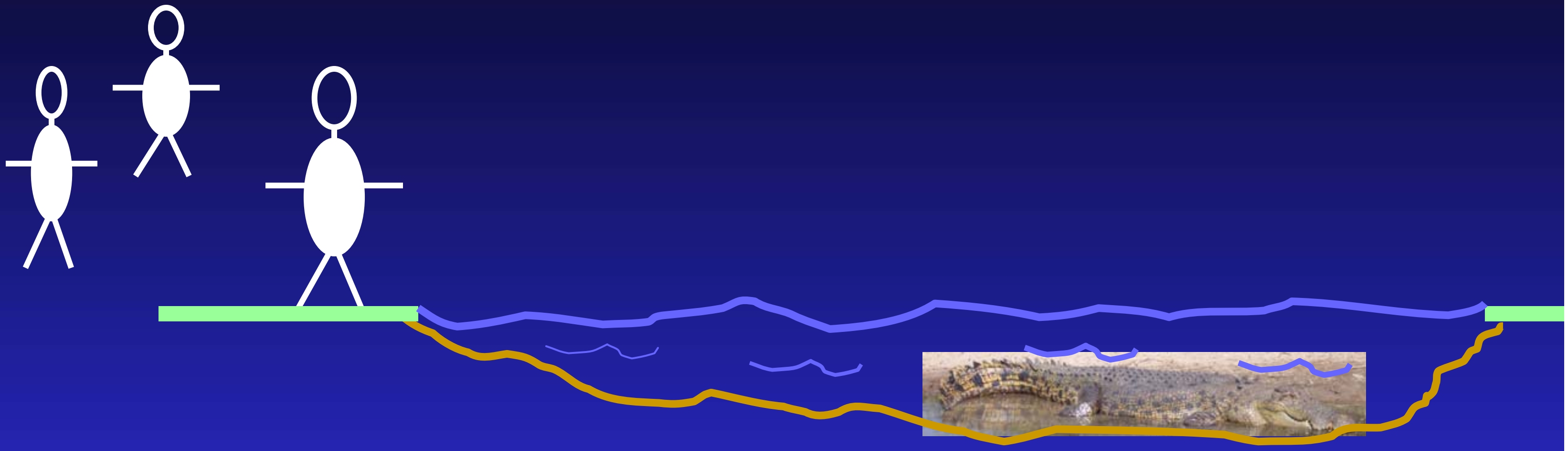


In the NAG board, the EPS corresponds to predicting with a collection (ensemble) of golf balls… but if reality is not a golf ball, then how do we interpret these distributions?

THAT IS ***THE*** QUESTION OF MODEL INADEQUACY:

How do we "verify" what a distribution of golf balls tell us about the single passage of the red ball?

Model Inadequacy and our three non-Floridian statisticians.

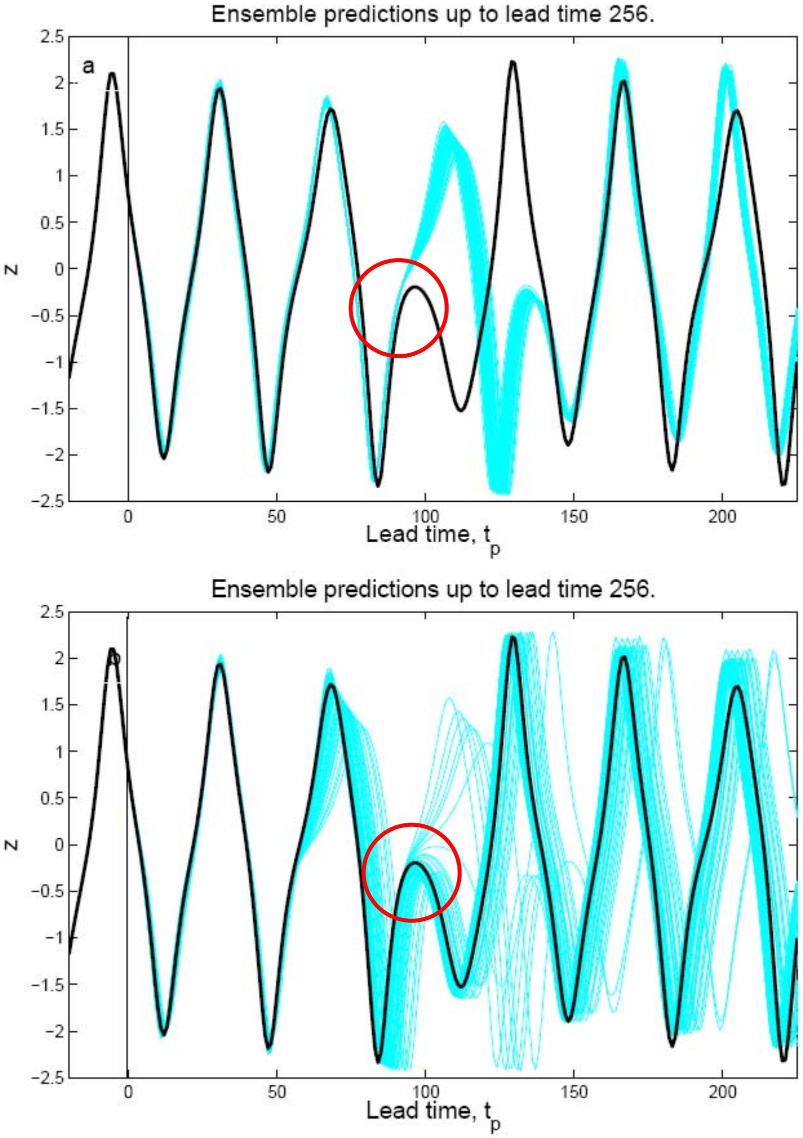


As it turns out, the river is rather shallow.

Model inadequacy covers things in the system that are not of the model.

The real question was could they make it across, the depth of the river was only one component…

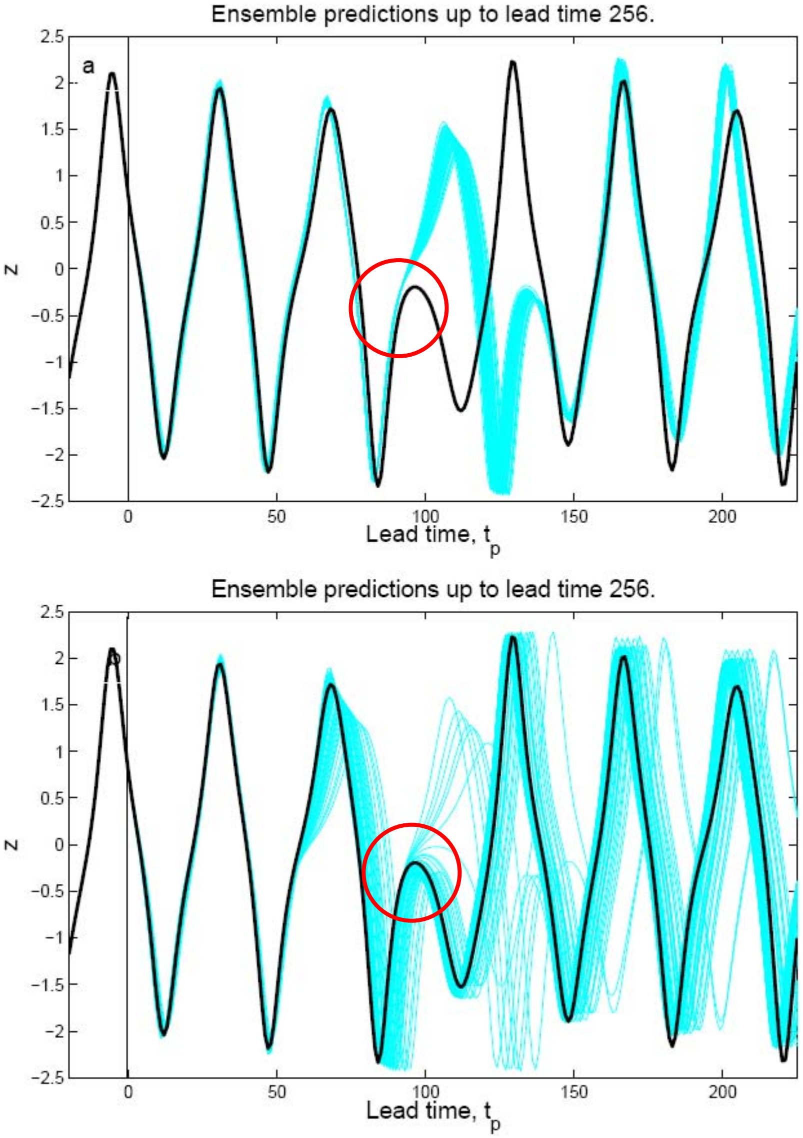
Moore-Spiegel Circuit (by Reason Machette)


Ensemble predictions up to lead time 256.

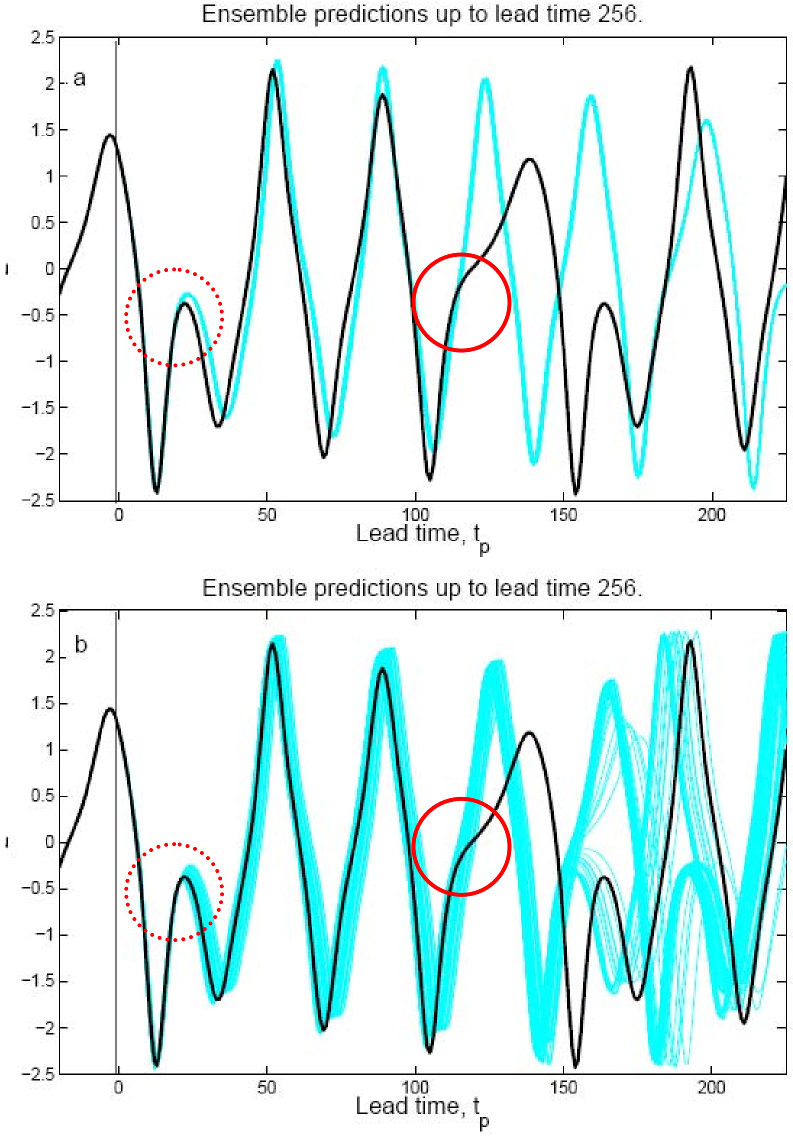
Ensemble predictions up to lead time 256.

Figure 7: Ensemble predictions using (a) model 1 and (b) model 2.

Moore-Spiegel Circuit (by Reason Machette)
What verification schemes can cope with this?



igure 7: Ensemble predictions using (a) model 1 and (b) model 2. The



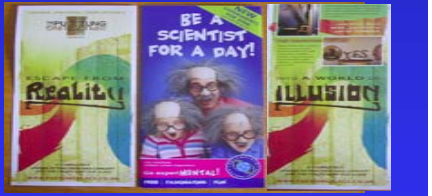
Ensemble predictions using (a) model 1 and (b) model 2

Mousie
(The best laid schemes of mice and men.)

Still thou art blest, compar'd wi me!
The present only toucheth thee:
But och! I backward cast my e'e,
On prospects drear!
An forward, tho I canna see,
I guess an fear!

Butterfly Effect: We do not know the initial conditions, the best we can hope for is a probability distribution; but this may require a (near) perfect model.

Is the goal of a forecast PDF reality or illusion?



Burns' Effect: We have to cope with model inadequacy: we *canna see* even a PDF: so how can we best *guess an fear*?

It depends on the task: we have to evaluate End-to-End.

When is a probabilistic Forecast not a probability forecast?

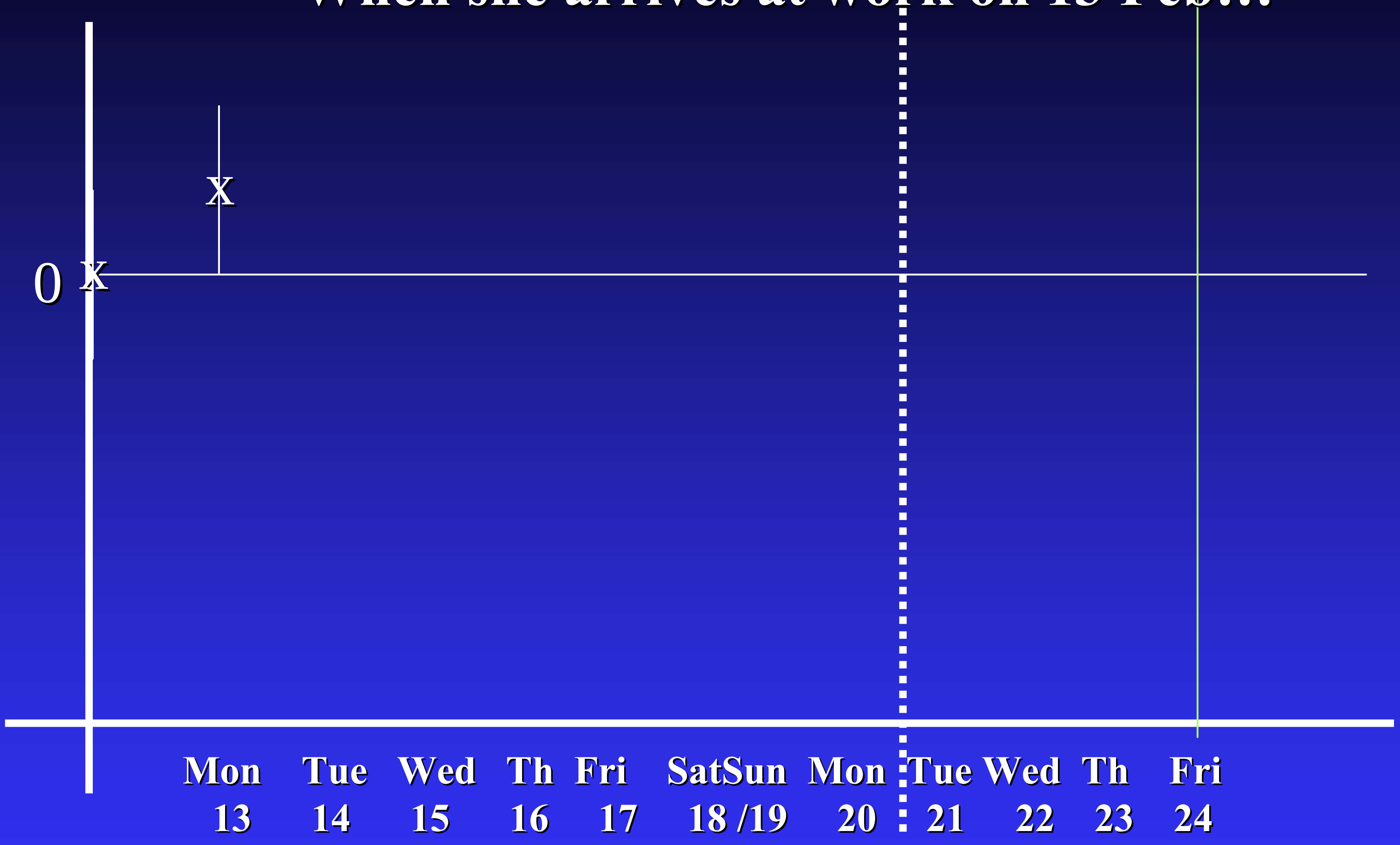
?Whenever you'd not apply it as a probability forecast?

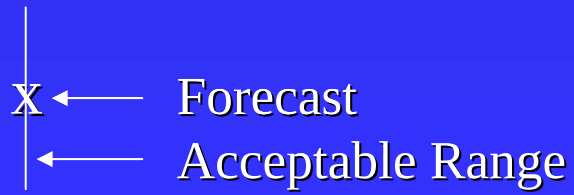
Numerate user's (Shell, Cal ISO, EDF, …) can detect that an operational forecast gives bad decision-support when used to maximise expected utility!

What can we do?
Say less, provide more:
A case study from the energy industry

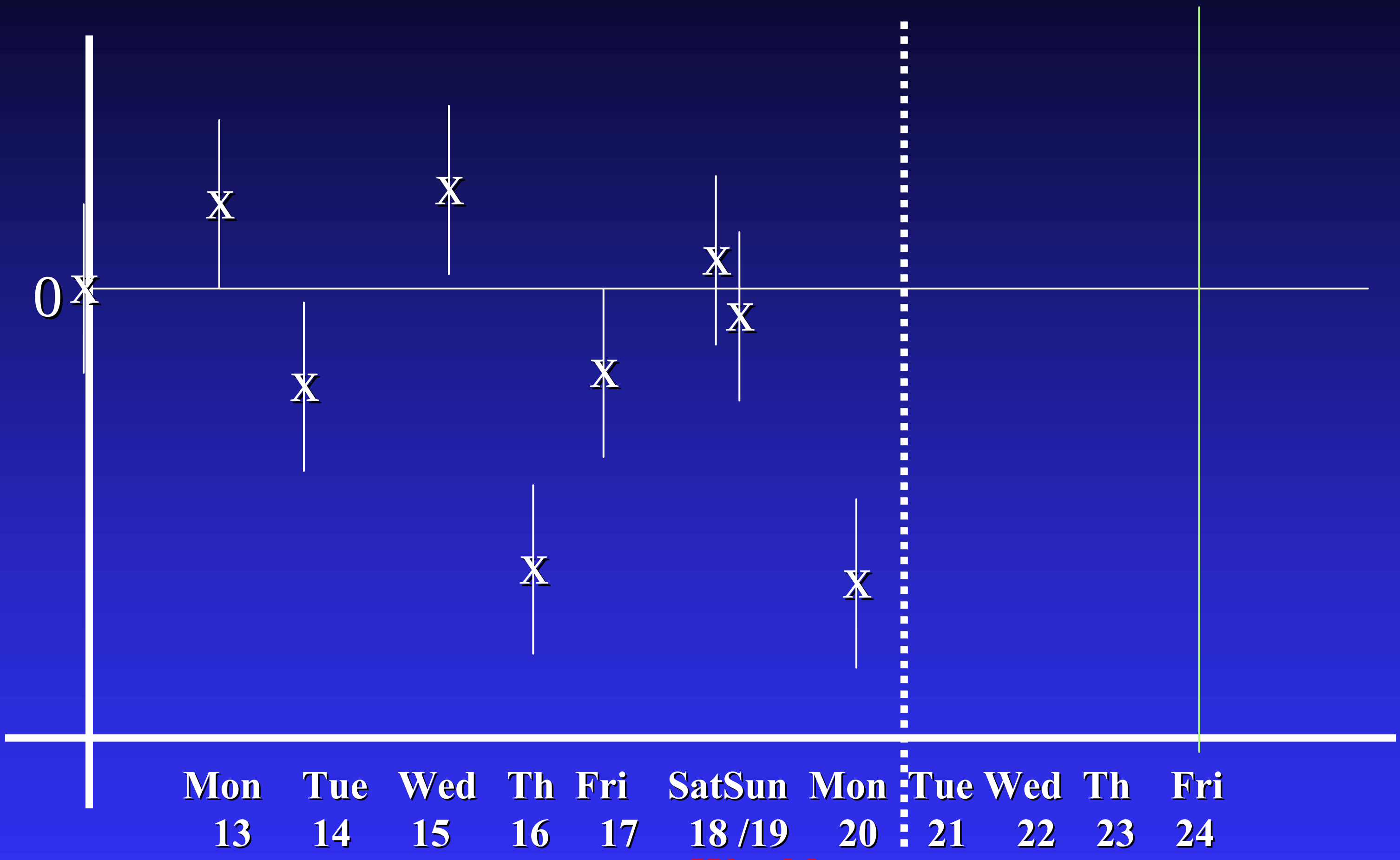
Charlize is buying gas to burn on 24 Feb
When she arrives at work on 13 Feb…

Forecast T_{eff} for Fri 24 Feb





Forecast T_{eff} for Fri 24 Feb



Would

Sell Buy Sell Buy Sell Have Buy

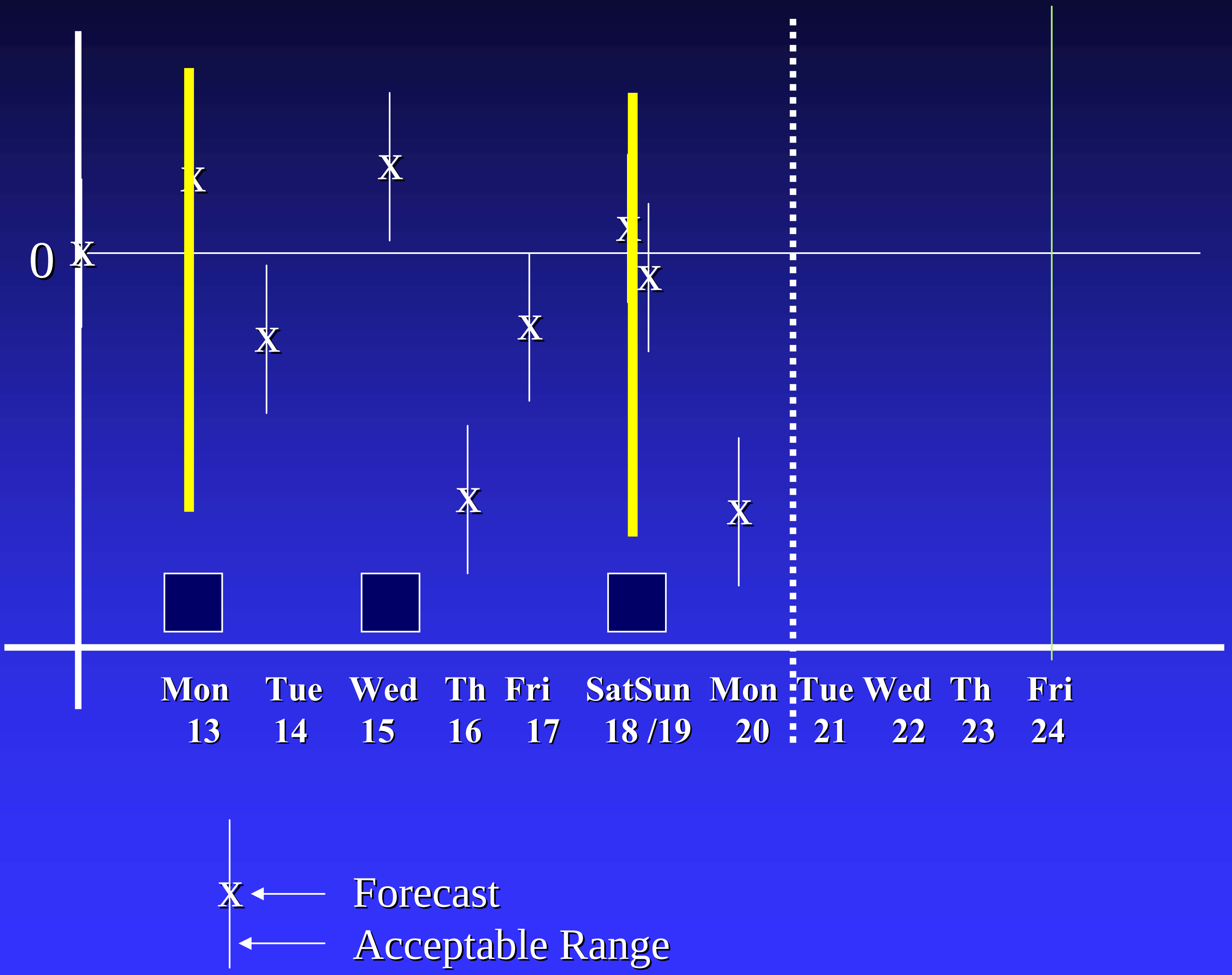
Sold

(but was home)

Note market would have jumped up here!

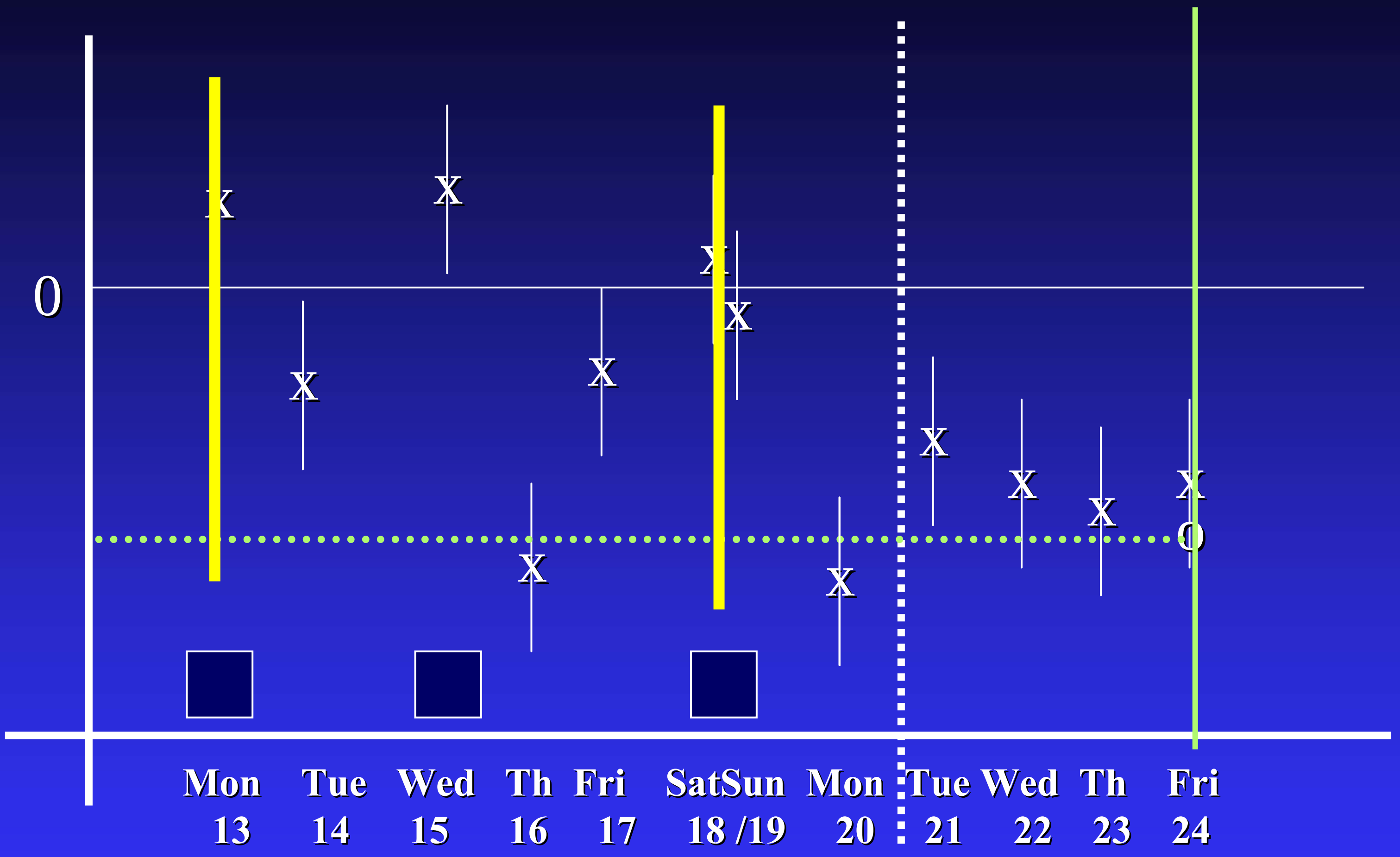
How can EPS probabilistic info help? And what happens???

Forecast T_{eff} for Fri 24 Feb

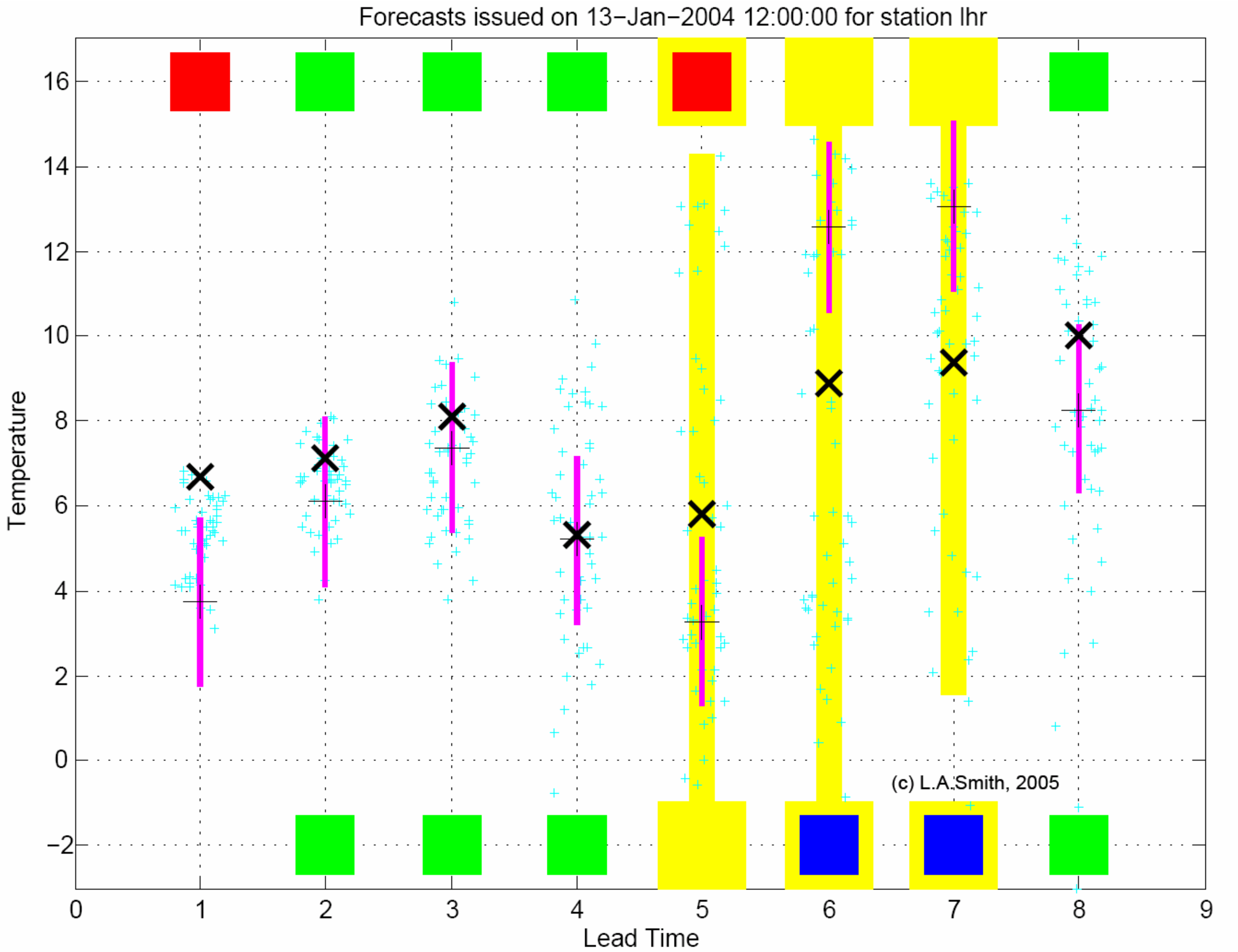


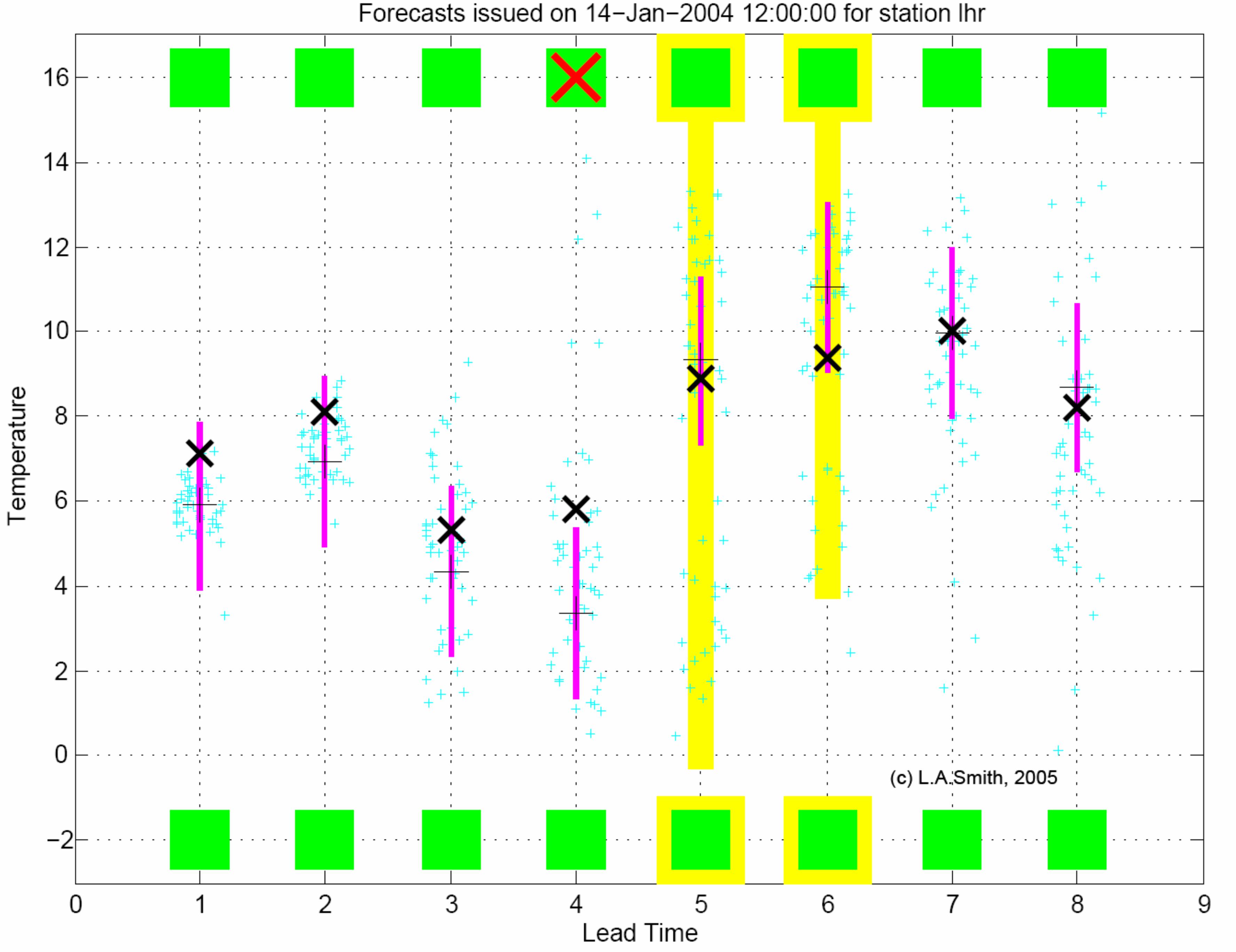
31 Jan 2007

Forecast T_{eff} for Fri 24 Feb

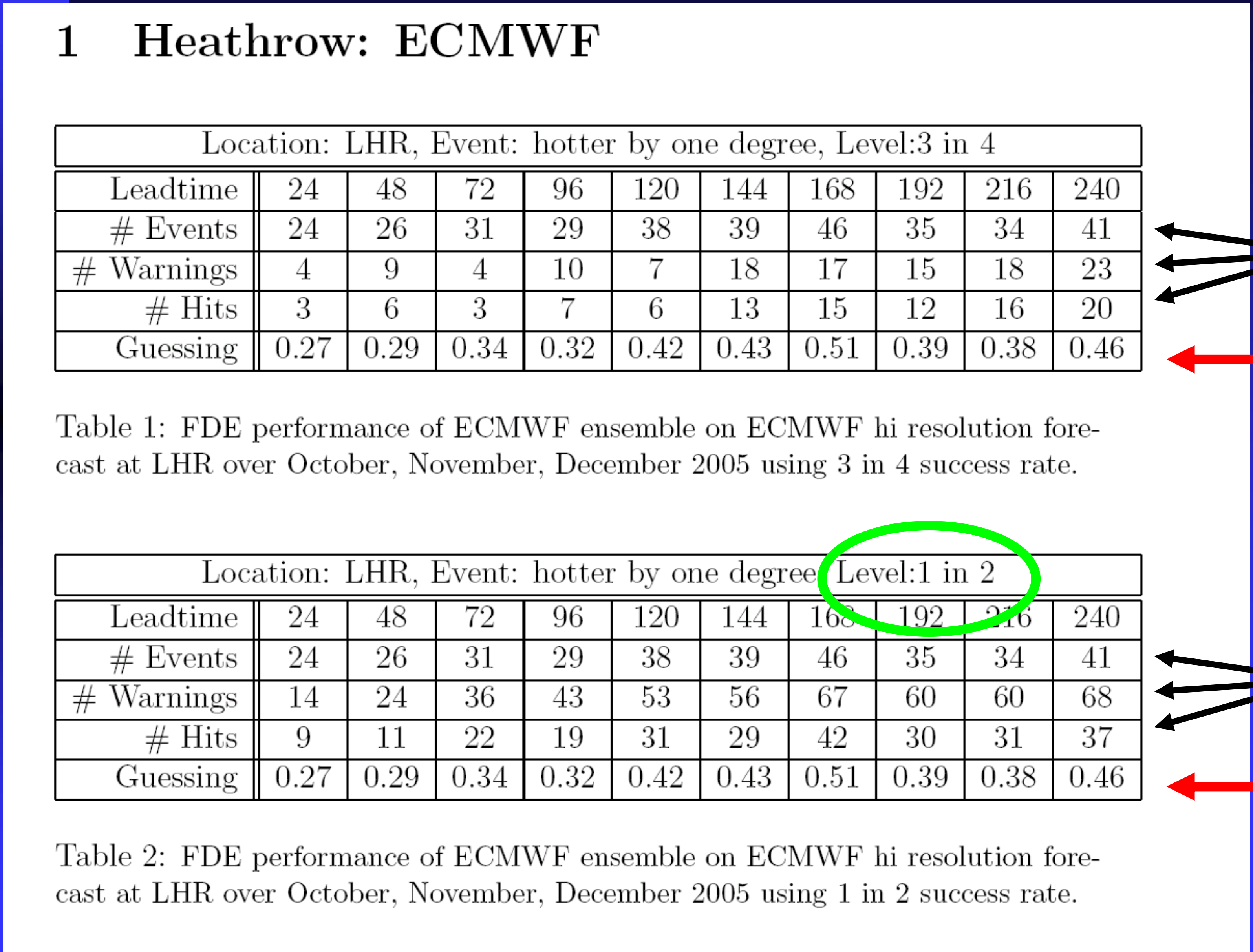


Probabilistic forecast guidance does not imply running an ensemble through “your” demand/generation/mix models!
But how would wqe verify this? (some effects are model error).


Forecasts issued on 13−Jan−2004 12:00:00 for station lhr
Temperature
16
14
12
10
8
6
4
2
0
−2
0
1
2
3
4
5
6
7
8
9
Lead Time
(c) L.A.Smith, 2005


Forecasts issued on 14−Jan−2004 12:00:00 for station lhr
Temperature
16
14
12
10
8
6
4
2
0
−2
Lead Time
0
1
2
3
4
5
6
7
8
9
(c) L.A.Smith, 2005

There are no “false alarms” here, just a user tolerance rate.
How can we best verify this?



1 Heathrow: ECMWF

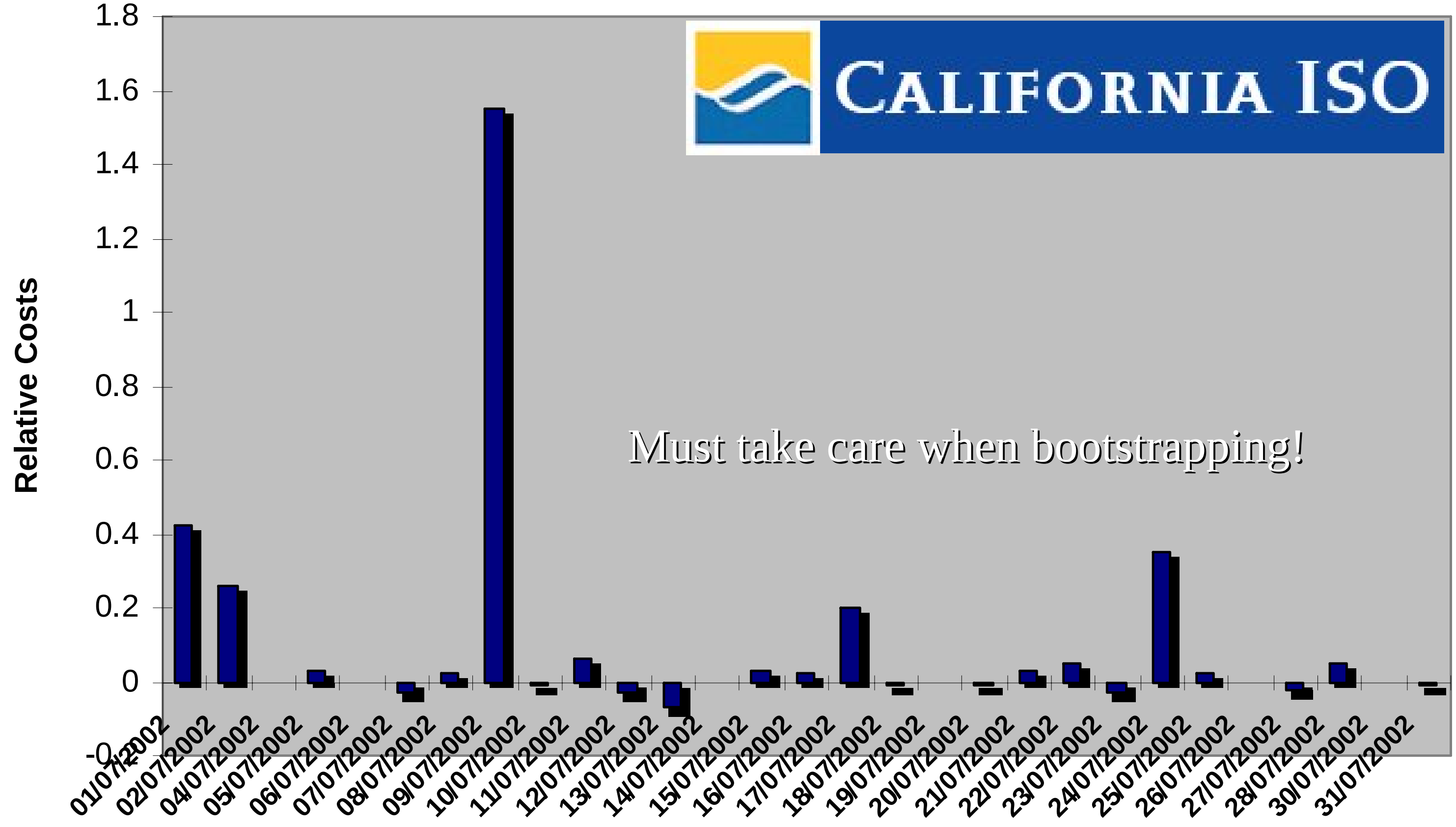
Location: LHR, Event: hotter by one degree, Level:3 in 4										
Leadtime	24	48	72	96	120	144	168	192	216	240
# Events	24	26	31	29	38	39	46	35	34	41
# Warnings	4	9	4	10	7	18	17	15	18	23
# Hits	3	6	3	7	6	13	15	12	16	20
Guessing	0.27	0.29	0.34	0.32	0.42	0.43	0.51	0.39	0.38	0.46

Table 1: FDE performance of ECMWF ensemble on ECMWF hi resolution forecast at LHR over October, November, December 2005 using 3 in 4 success rate.

Location: LHR, Event: hotter by one degree, Level:1 in 2										
Leadtime	24	48	72	96	120	144	168	192	216	240
# Events	24	26	31	29	38	39	46	35	34	41
# Warnings	14	24	36	43	53	56	67	60	60	68
# Hits	9	11	22	19	31	29	42	30	31	37
Guessing	0.27	0.29	0.34	0.32	0.42	0.43	0.51	0.39	0.38	0.46

Table 2: FDE performance of ECMWF ensemble on ECMWF hi resolution forecast at LHR over October, November, December 2005 using 1 in 2 success rate.

In this case, interpreting the ensemble as a probability (and maximizing Expected Utility) is far from optimal: is it then rational to interpret the forecast distribution as a probability forecast?



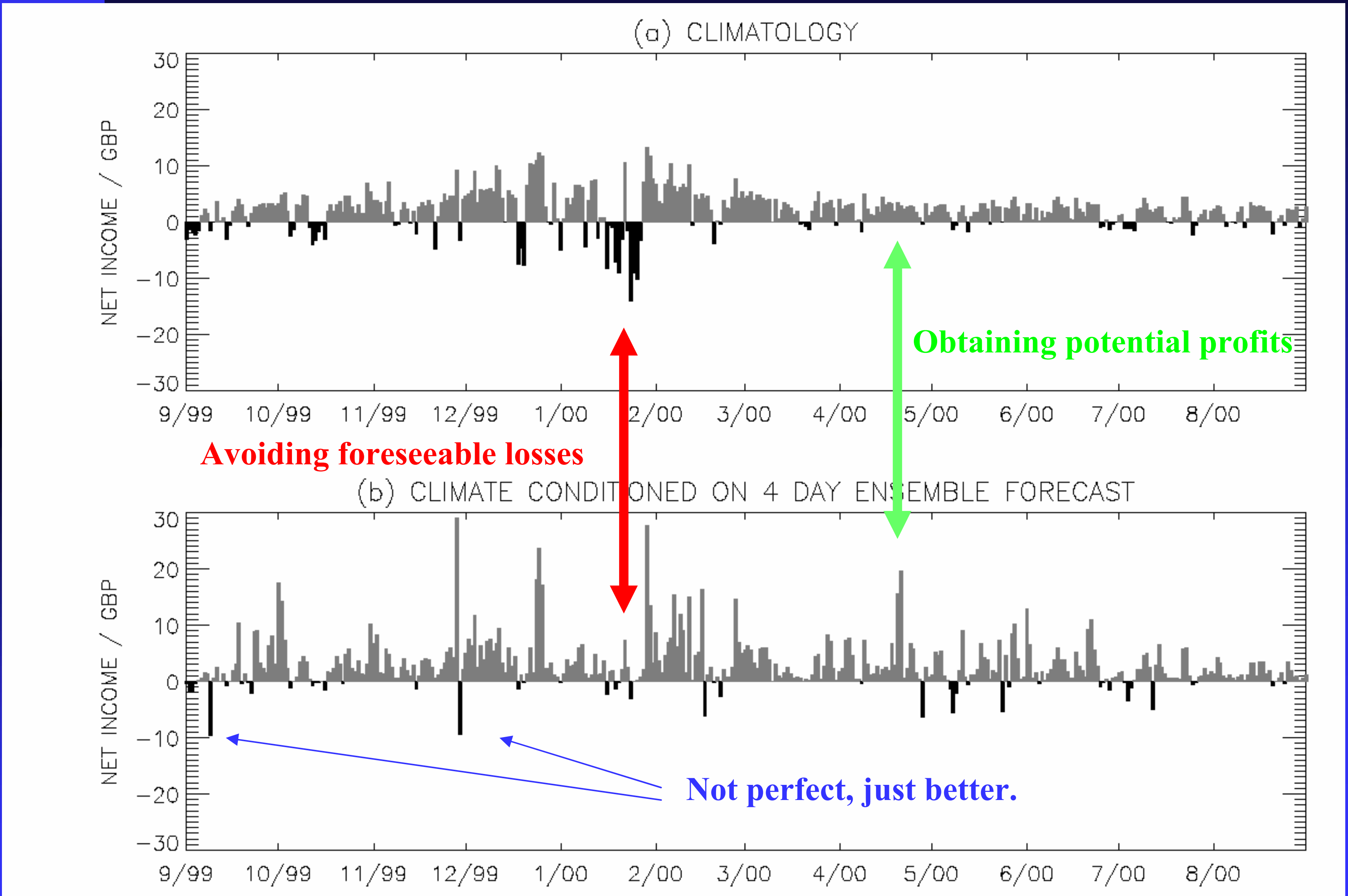
Must take care when bootstrapping!

From Smith, Altalo & Ziehmann (2004)

Figure 6: Relative costs of PF1 forecasts versus the Cal ISO surrogate forecasts for days in July 2002, a positive value represents a savings of using PF1. Note the significant savings on July 9th.

Wind farm profits

Roulston & Smith (2003)

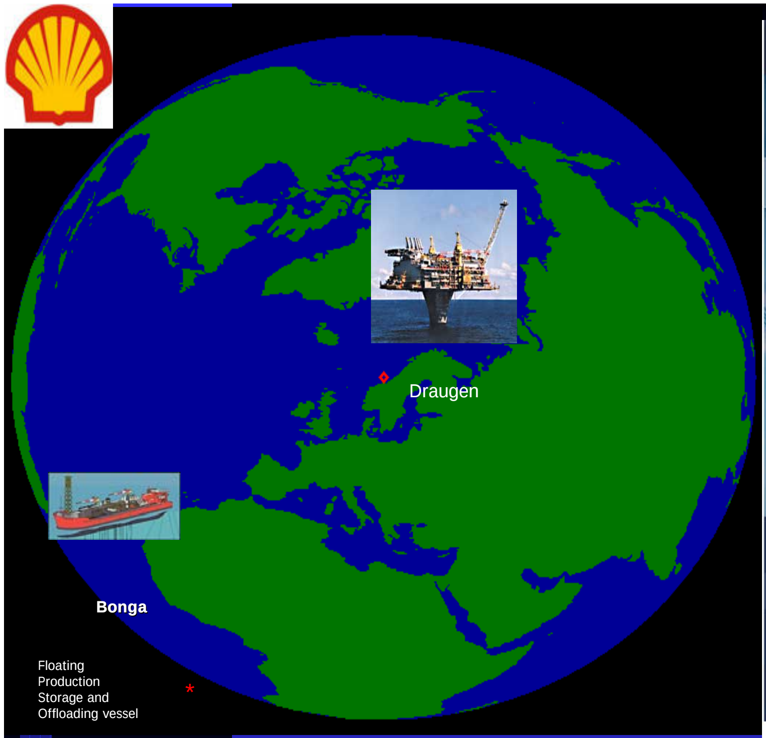


Obtaining potential profits

Avoiding foreseeable losses

Not perfect, just better.

Any empirically meaningful user metric!



SCIENCE APPLICATIONS INTERNATIONAL CORPORATION

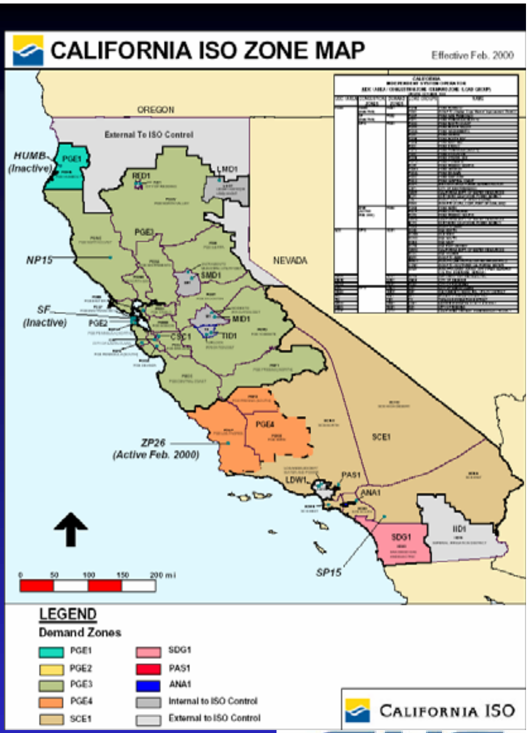
The Economic Benefit of Incorporating Weather and Climate Forecasts Into Western Energy Production Management.

Deliverable 2: **Case Study Design and Scenario Report**

Scripps Institution of Oceanography (SIO)
LA Jolla, CA

December 31, 2003

SAIC
An Employee-Owned Company



Scottish and Southern Energy plc

National Grid

SAIC

Seeing Through Our Models

(Empirically based decision support in the weather scenario: Any PDFs?)

CALIFORNIA ISO

METRA
METRA ENERGY

ALSTOM

LONDON ELECTRICITY

THE WEATHER CHANNEL
weather.com

DEPARTMENT OF THE NAVY
ONR
Science & Technology

EDF ENERGY

British Energy

Risk Management Solutions

RMS™

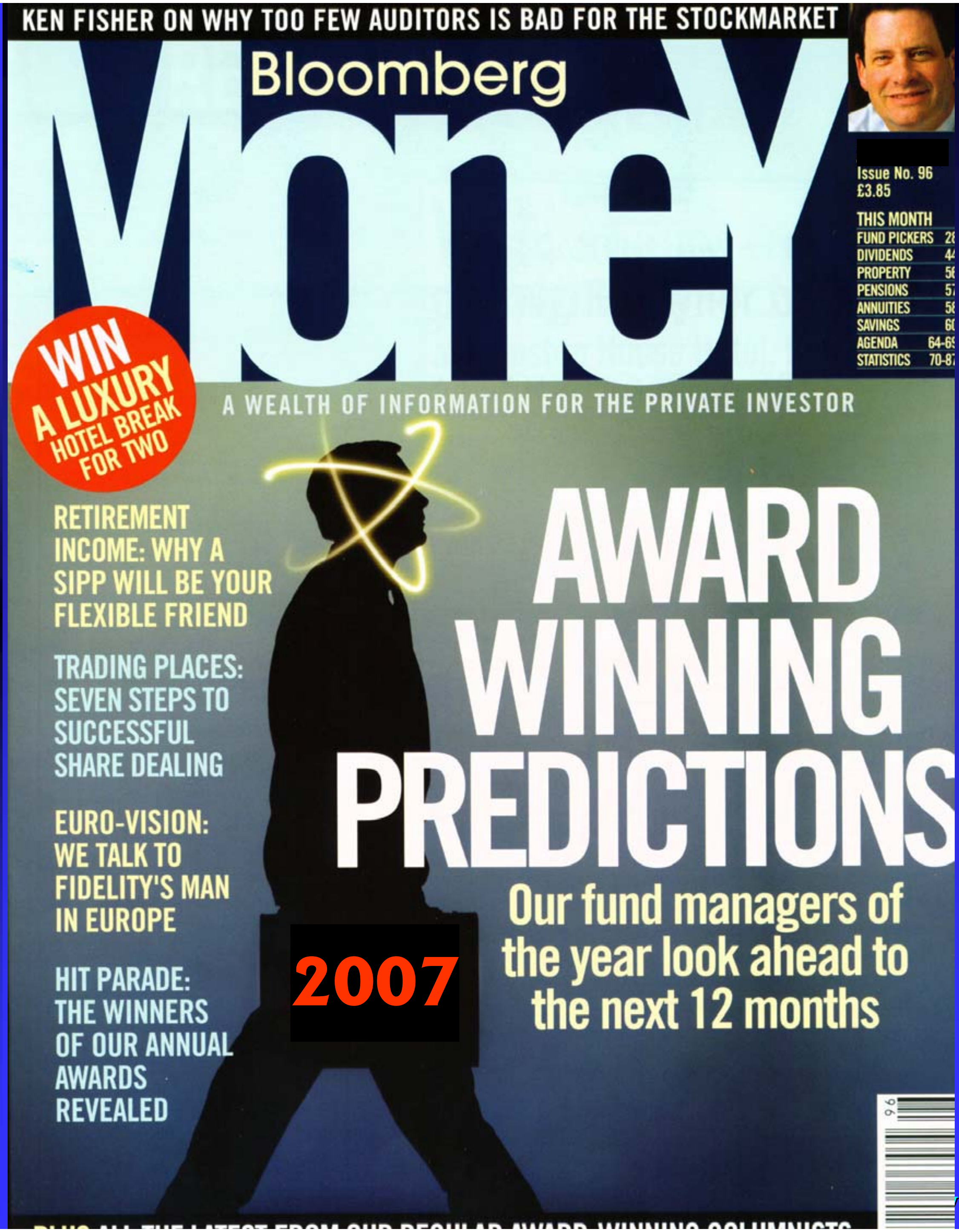
Conclusions

Verification Methods aim model improvement, forecast improvement, and ideally, quantify the utility for the user.

There are many methods, and certain properties should be respected (like using proper scores).

Relevance, however, is as important as *rigor;* new approaches should be developed as needed.

Uncertainty should always be estimated and illustrated.



KEN FISHER ON WHY TOO FEW AUDITORS IS BAD FOR THE STOCKMARKET

Bloomberg Money

Issue No. 96
£3.85

THIS MONTH

FUND PICKERS	28
DIVIDENDS	44
PROPERTY	5
PENSIONS	5
ANNUITIES	58
SAVINGS	6
AGENDA	64-6
STATISTICS	70-8

WIN A LUXURY HOTEL BREAK FOR TWO

A WEALTH OF INFORMATION FOR THE PRIVATE INVESTOR

RETIREMENT INCOME: WHY A SIPP WILL BE YOUR FLEXIBLE FRIEND

TRADING PLACES: SEVEN STEPS TO SUCCESSFUL SHARE DEALING

EURO-VISION: WE TALK TO FIDELITY'S MAN IN EUROPE

HIT PARADE: THE WINNERS OF OUR ANNUAL AWARDS REVEALED

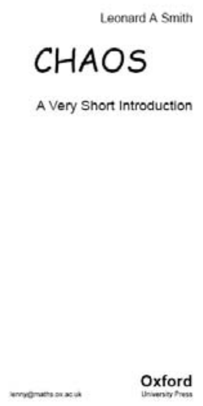
AWARD WINNING PREDICTIONS

Our fund managers of the year look ahead to the next 12 months

2007

PLUS ALL THE LATEST FROM OUR REGULAR AWARD-WINNING COLUMNISTS

When should a prediction win an award?


Leonard A Smith

CHAOS

A Very Short Introduction

Oxford University Press

lenny@maths.ox.ac.uk

Additional References

M.S. Roulston D.T. Kaplan, J. Hardenberg & L.A. Smith (2003) Using Medium Range Weather Forecasts to Improve the Value of Wind Energy Production. *Renewable Energy* **28** (4) 585–602

LA Smith (2003) Predictability Past Predictability Present. ECMWF Seminar on Predictability. NOW in a CUP book (ed. Tim Palmer).

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

Broecker & Smith (2007a,b,c...) www.lsecats.org

Nancy Cartwright (1983) *How the Laws of Physics Lie*, OUP

www.lsecats.org

lenny@maths.ox.ac.uk

Live Discussion Board in Now