# Real Time Series analysis and modelling

## Aid machine learning with dynamical insight

**Hailiang Du**
Department of Mathematical Sciences, Durham University
hailiang.du@durham.ac.uk
Data Science Institute, London School of Economics and Political Science
h.l.du@lse.ac.uk

All theorems are true, All models are wrong. All data are inaccurate. What are we to do?

The aim of this course is to teach you how to deal with real data, to increase your scepticism regarding reliable modelling in practice, and to expand the tool box you carry to include nonlinear techniques, both deterministic and stochastic with the aid of dynamical insight.

In short: to get you to think before you compute (and perhaps afterwards too.)

# Lecture 6

Forecast interpretation and evaluation

# How did you evaluate forecasts?

1. Root Mean Square Error
2. Correlation
3. Visualisation
4. Rank Histogram (Talagrand diagram)
5. Brier score
6. Reliability Diagram
7. Rank Probability Score
8. Proper Linear Score
9. Ignorance (log p score)
10. Others…

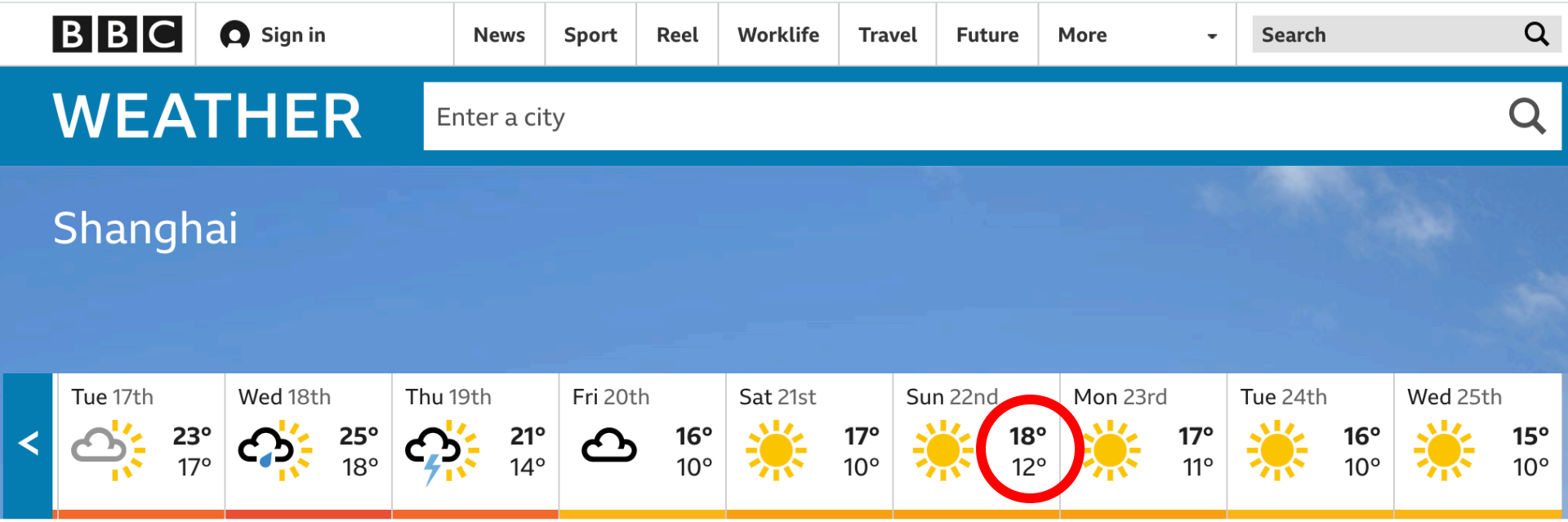
# Point (deterministic) Forecasts

Forecast: Head

Given a fair coin and a fair toss, my point forecast have 50% chance to be right and 50% chance to be wrong.

Discrete variable
Finite number of outcomes

# Point (deterministic) Forecasts




**Forecast:**
Maximum Temperature on Sunday is 18 degrees

This kind of point (deterministic) forecast is almost always wrong.

Continuous variable
Infinite number of outcomes

# Uncertainties

- Initial condition uncertainty
- Parameter uncertainty
- Model structure uncertainty

Point forecast → Deterministic forecast
is unable to explore any of the uncertainties.

# Ensemble forecasts

- Initial condition ensemble

  explore initial condition uncertainty

- Parameter ensemble

  explore parameter uncertainty

- Multi-model ensemble

  explore model structure uncertainty

# (Initial condition) Ensemble forecasts

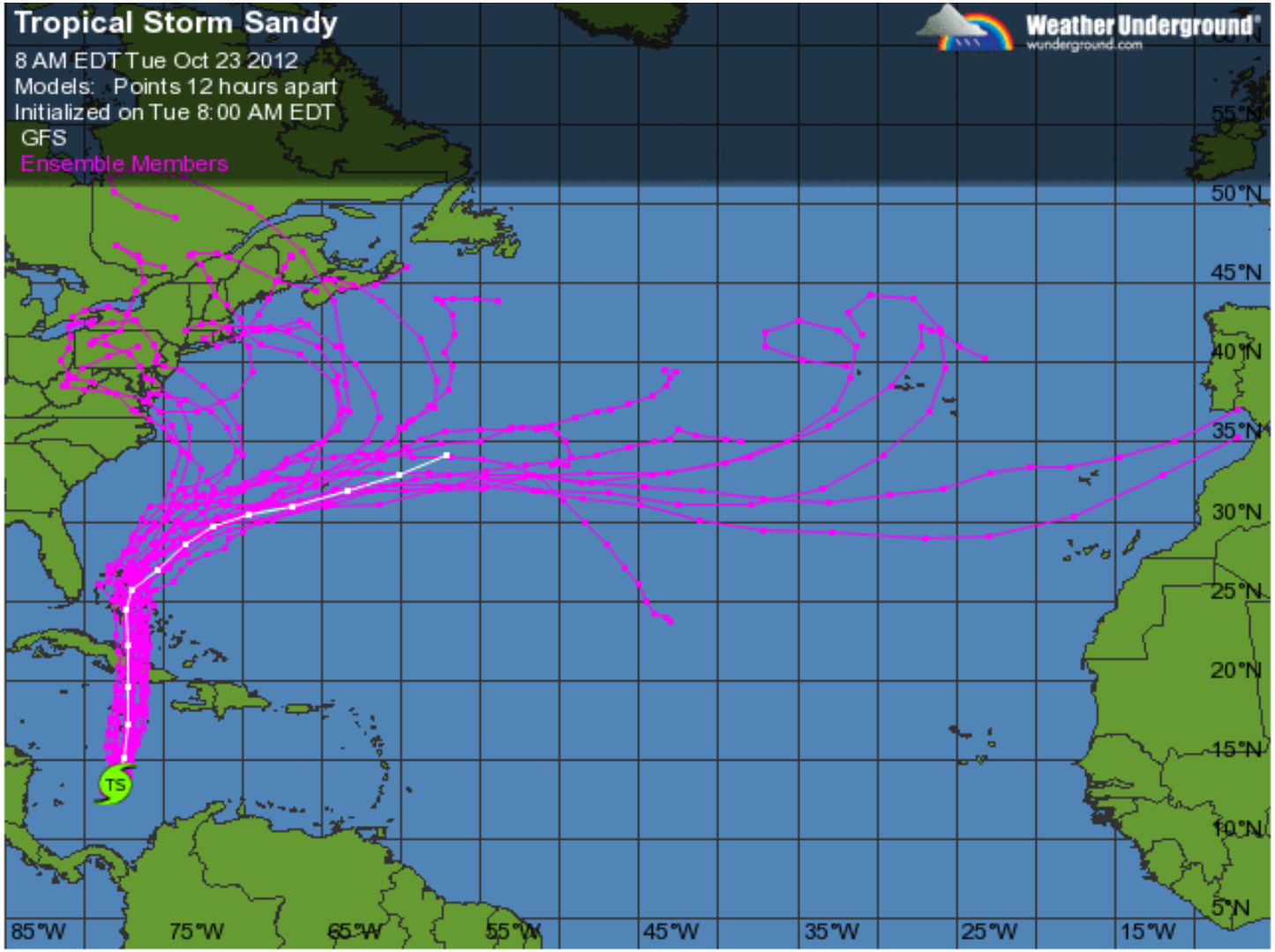
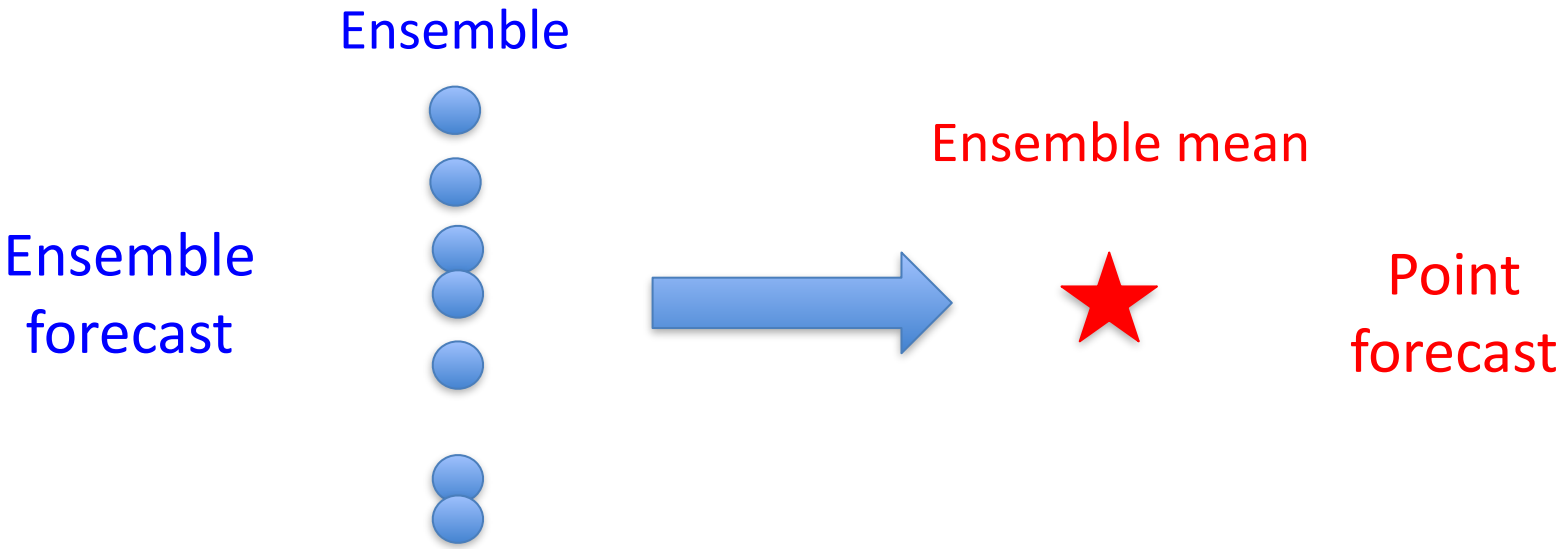


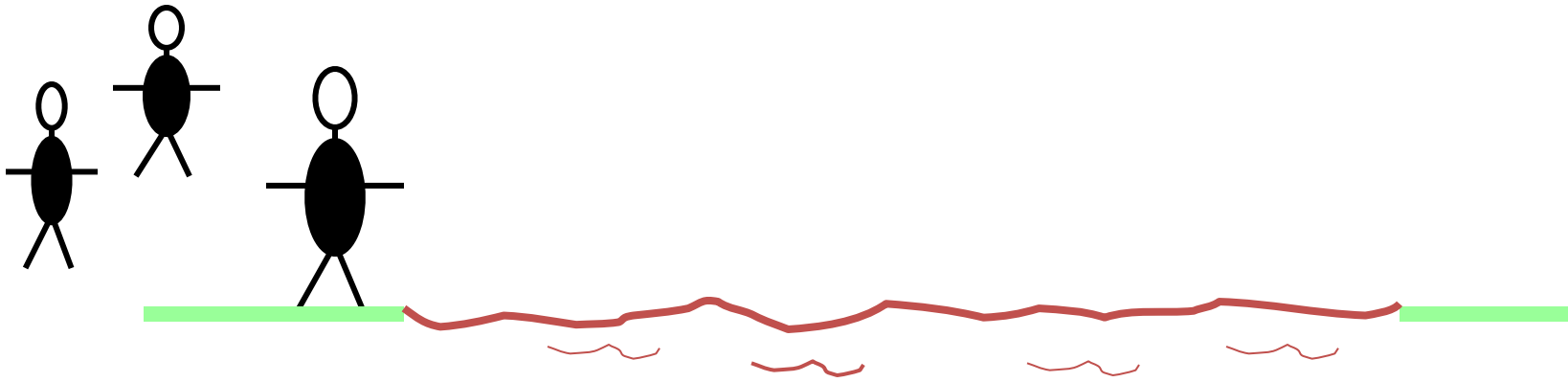

**Figure 2.** The Tuesday morning 06Z (2 am EDT) run of the GFS model was done 20 times at lower resolution with slightly varying initial conditions of temperature, pressure, and moisture to generate an ensemble of forecast tracks for Sandy (pink lines). These forecasts show substantial uncertainty in Sandy's path after Friday, with the majority of the forecasts taking Sandy to the northeast, out to sea, but a substantial number predicting a landfall in the Northeast or mid-Atlantic states of the U.S. The white line shows the official GFS forecast, run at higher resolution.

# Interpret ensemble forecast using the ensemble mean




"It is easy to calculate the ensemble mean."
"It is easy to evaluate the ensemble mean forecast."
"The ensemble mean contains "most" of information of the ensemble"
"Industry standard"

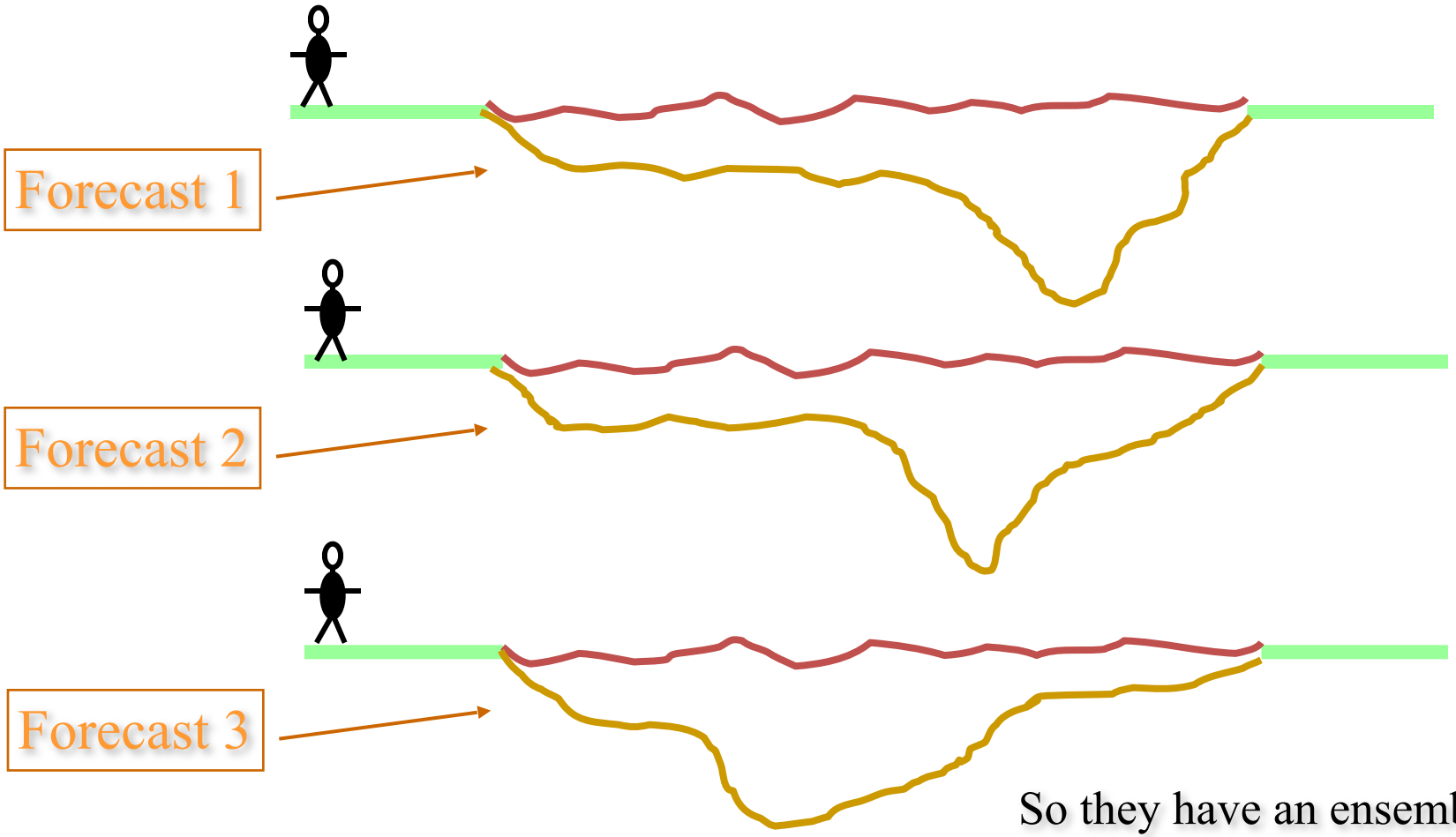
# The Parable of the Three Statisticians.



Three Statisticians come to a river, they want to know if they can cross safely.
(They cannot swim.)

# Three statisticians wish to cross a river

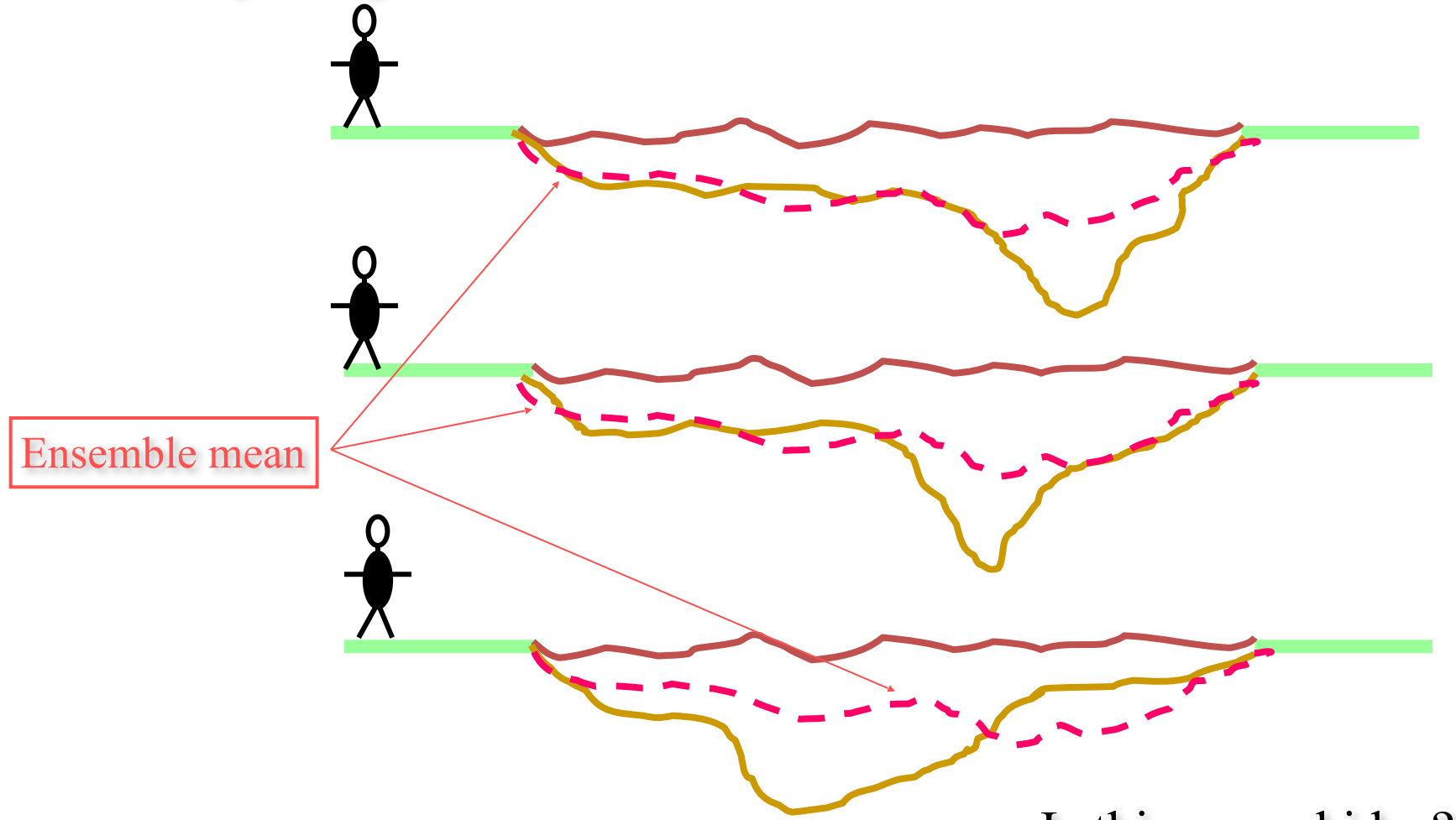
Each has a forecast of depth which indicates they will drown.




So they have an ensemble forecast,with three members

## Three 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…




Is this a good idea?

# Ensemble mean would have destroyed this informative forecast

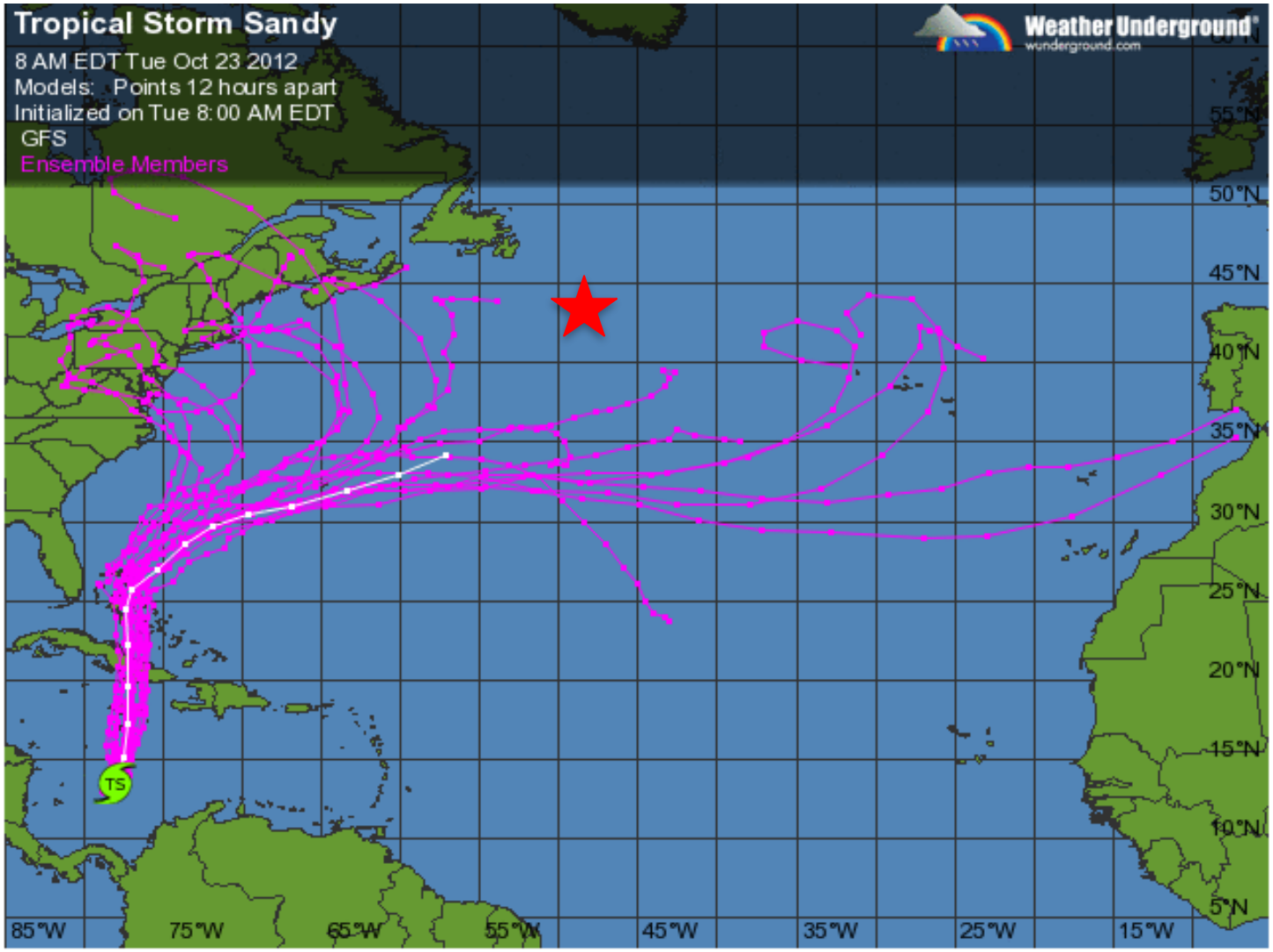
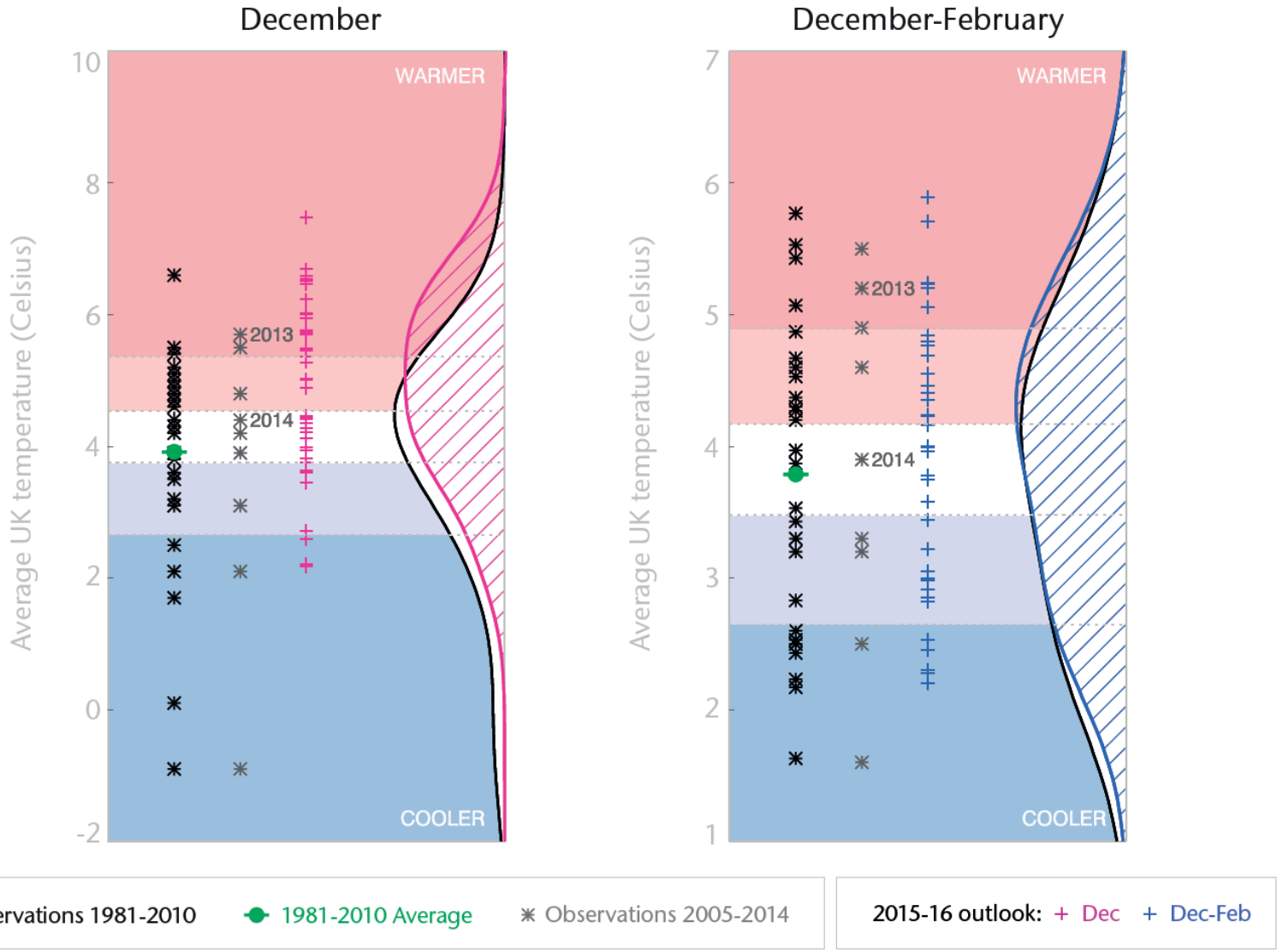



**Figure 2.** The Tuesday morning 06Z (2 am EDT) run of the GFS model was done 20 times at lower resolution with slightly varying initial conditions of temperature, pressure, and moisture to generate an ensemble of forecast tracks for Sandy (pink lines). These forecasts show substantial uncertainty in Sandy's path after Friday, with the majority of the forecasts taking Sandy to the northeast, out to sea, but a substantial number predicting a landfall in the Northeast or mid-Atlantic states of the U.S. The white line shows the official GFS forecast, run at higher resolution.

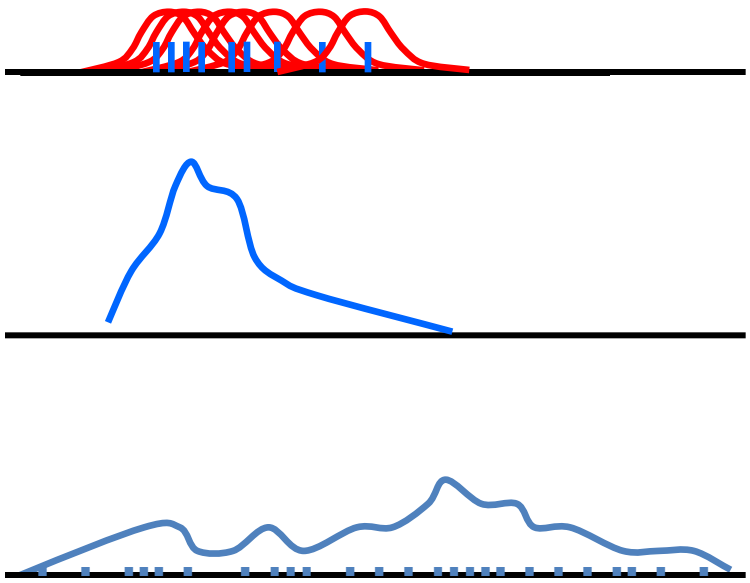
# From Ensembles to Predictive Distributions

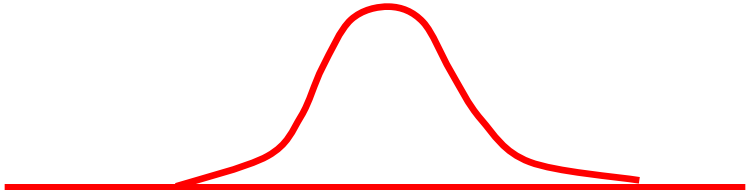
1-month and 3-month UK outlook for temperature in the context of observed climatology



# Ensembles Members In - Predictive Distributions Out

K is the kernel, with parameter σ for example





$$P(y) = \sum_{i=1}^{N_{ens}} K(y, x_i)/N_{ens}$$

$$P_{clim}(y) = \sum_{i=1}^{N_{clim}} K(y, s_i)/N_{clim}$$

B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1998.

J Bröcker, LA Smith (2008) From Ensemble Forecasts to Predictive Distribution Functions *Tellus A* 60(4): 663.

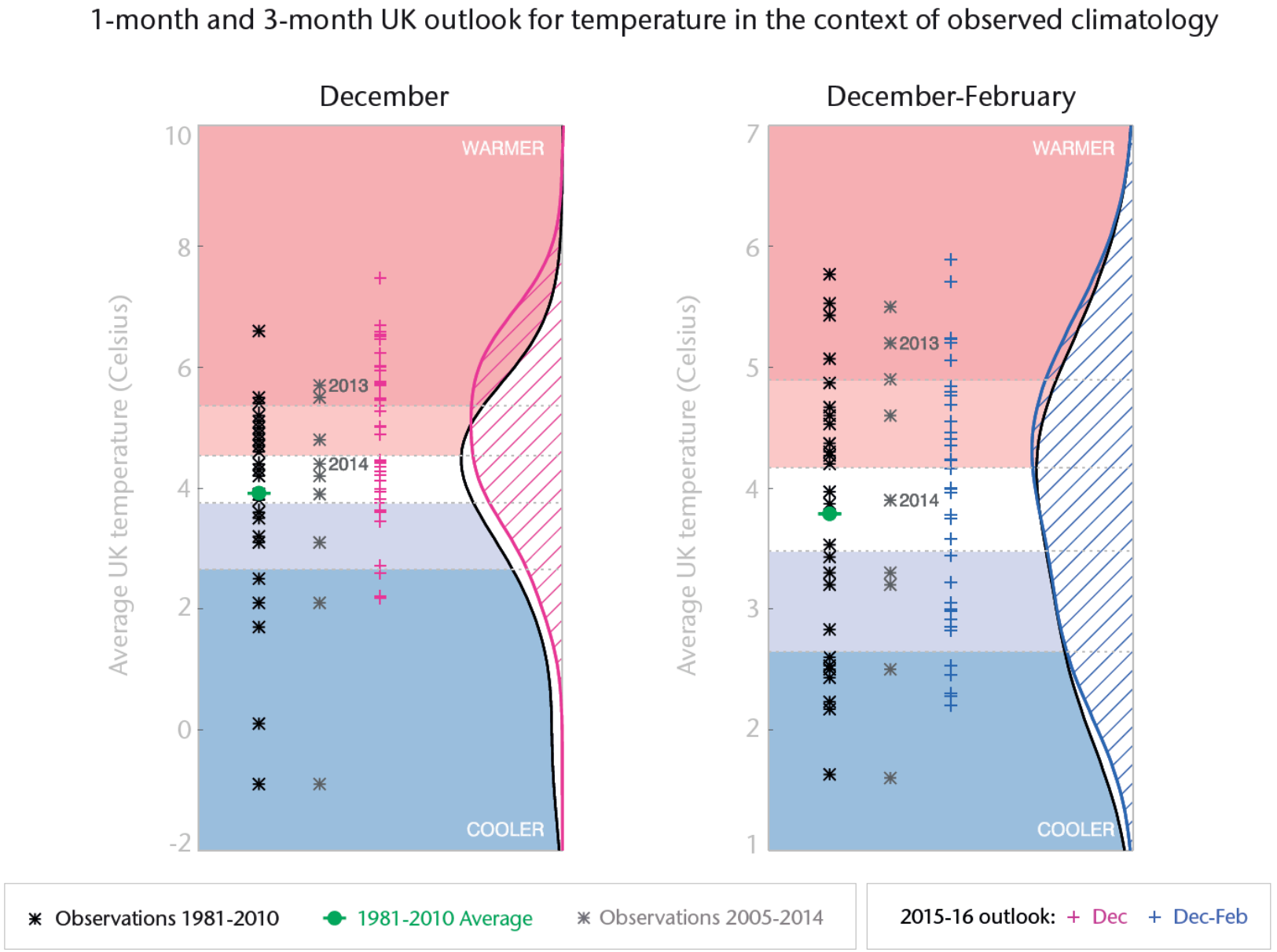
# Probabilistic Forecasts

Forecast: 42% chance Head; 58% chance Tail.

Given any finite number of toss, my probabilistic forecast can not been shown to be wrong.

Hypothesis test?

# From Ensembles to Predictive Distributions




From http://search.metoffice.gov.uk/kb5/metoffice/metoffice/results.page?qt=three+month+outlook Not there any more

# Evaluate probabilistic forecast

2009 July

The aim is not to shown whether the forecasts are right or wrong, but to shown which forecast system is better, and how much better.


## MET OFFICE'S CHANGING TUNE

**APRIL**

"The coming summer is odds on for a ... to be warmer than average and rainfall near or below average for the three months of summer."

**YESTERDAY** July

"In April we said there was a 65% chance of temperatures above average and rainfall below average but that does leave a 35% chance that the opposite would be true. But seasonal forecasting is still a new science."

**Probabilities will be seen as a get-out clause** *(Meteorologists are just covering their backs)*

# Diagnostic tools

From IPCC (Intergovernmental Panel on Climate Change) report Page 755:

Annan and Hargreaves (2010) have proposed a 'rank histogram' approach to evaluate model ensembles as a whole, rather than individual models, by diagnosing whether observations can be considered statistically indistinguishable from a model ensemble. Studies based on this approach have suggested that MMEs (CMIP3/5) are 'reliable' in that they are not too narrow or too dispersive as a sample of possible models, but existing single-model-based ensembles tend to be too narrow (Yokohata et al., 2012, 2013).

# Diagnostic tools

From IPCC (Intergovernmental Panel on Climate Change) report Page 755:

Annan and Hargreaves (2010) have proposed a 'rank histogram' approach to evaluate model ensembles as a whole, rather than individual models, by diagnosing whether observations can be considered statistically indistinguishable from a model ensemble. Studies based on this approach have suggested that MMEs (CMIP3/5) are 'reliable' in that they are not too narrow or too dispersive as a sample of possible models, but existing single-model-based ensembles tend to be too narrow (Yokohata et al., 2012, 2013).

Anderson, J. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Clim., 9(7), 1518 – 1530

Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction systems, paper presented at the Workshop on Predictability, Eur. Cent. for Medium-Range Weather Forecasts, Reading, U. K. 1977.

# Diagnostic tools

- Rank Histogram (Talagrand Diagram)

  ➢The assumption to be tested is that truth is drawn from the same distribution as the ensemble members.

  ➢If this is true then the observed verification is equally likely to fall between any two ensemble members or outside the ensemble.

  ➢This can be quantified by constructing a histogram that counts the number of times the verification falls between each ensemble member.

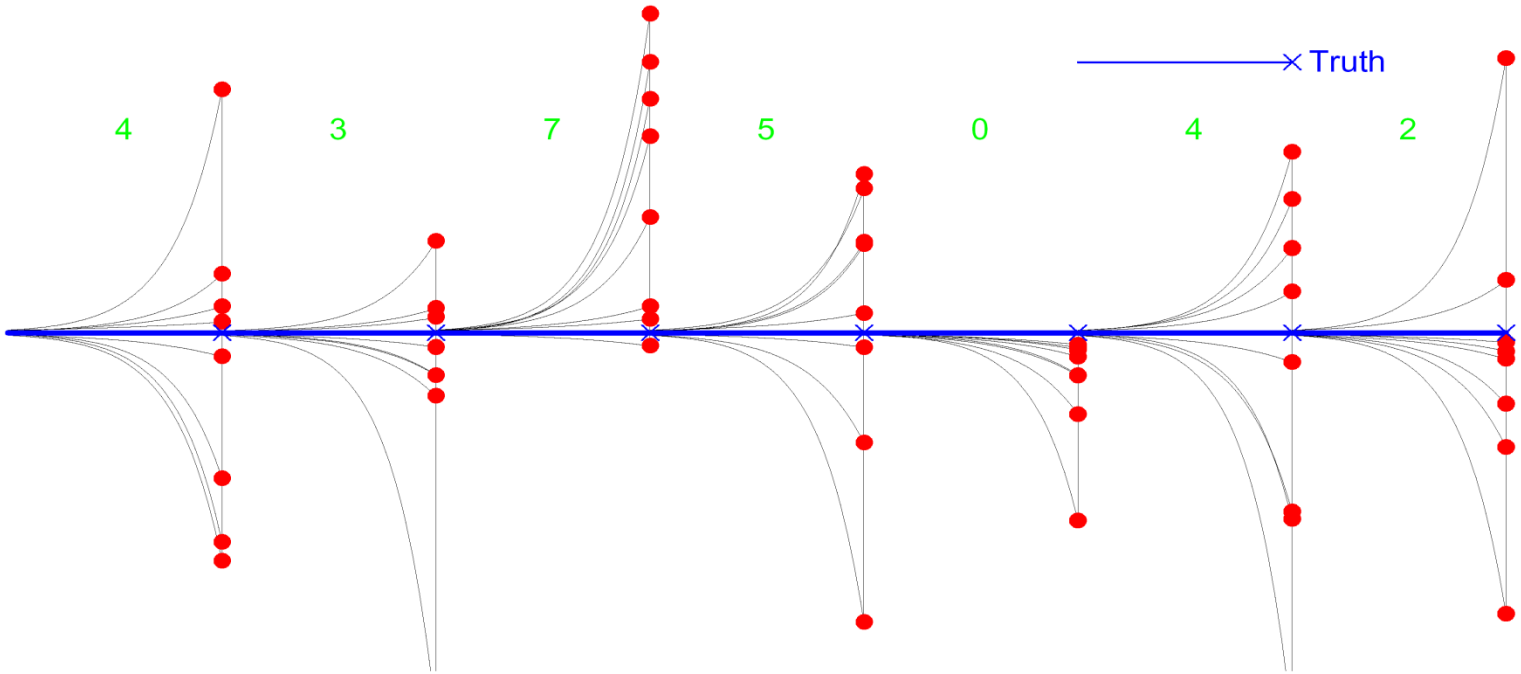
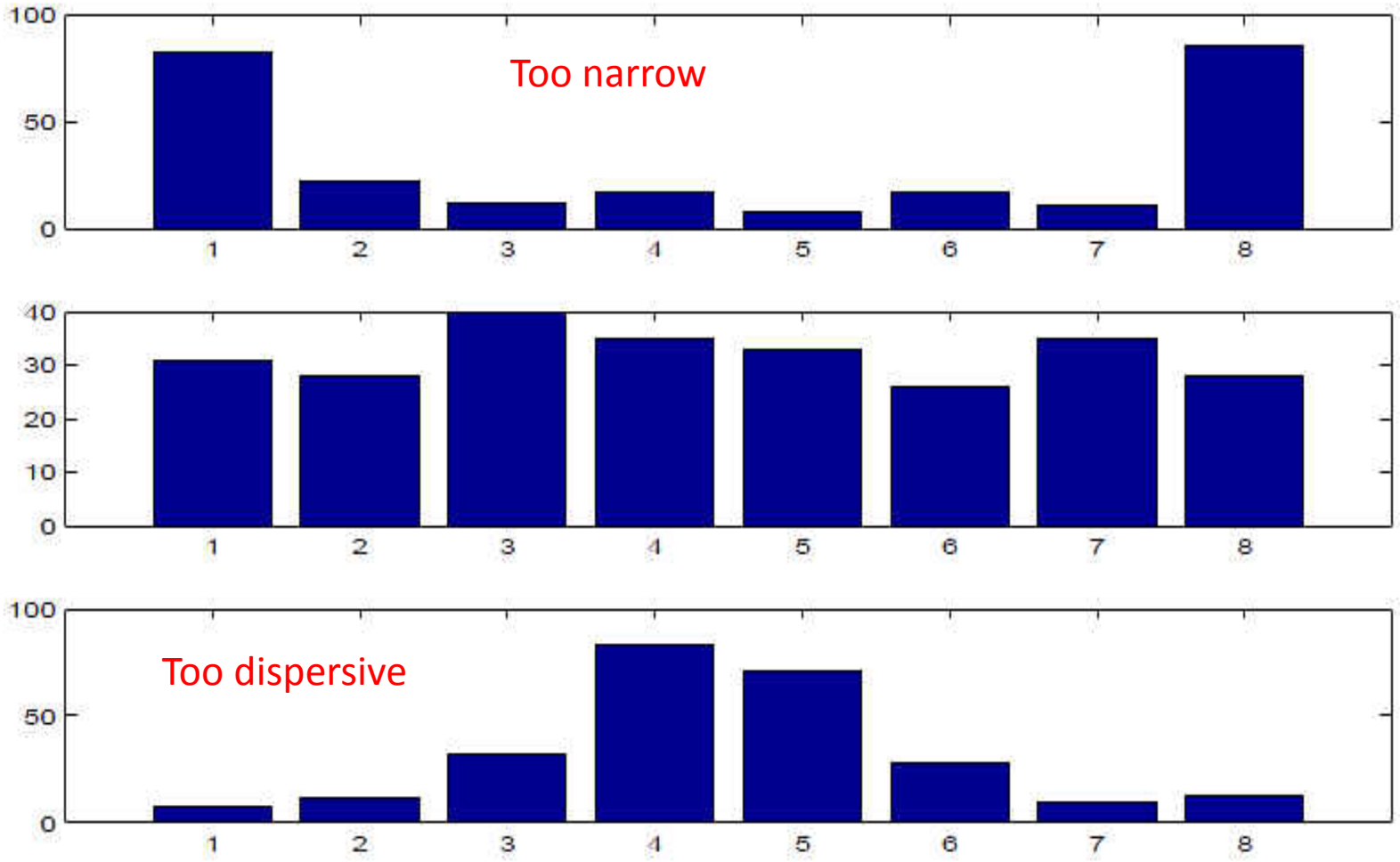


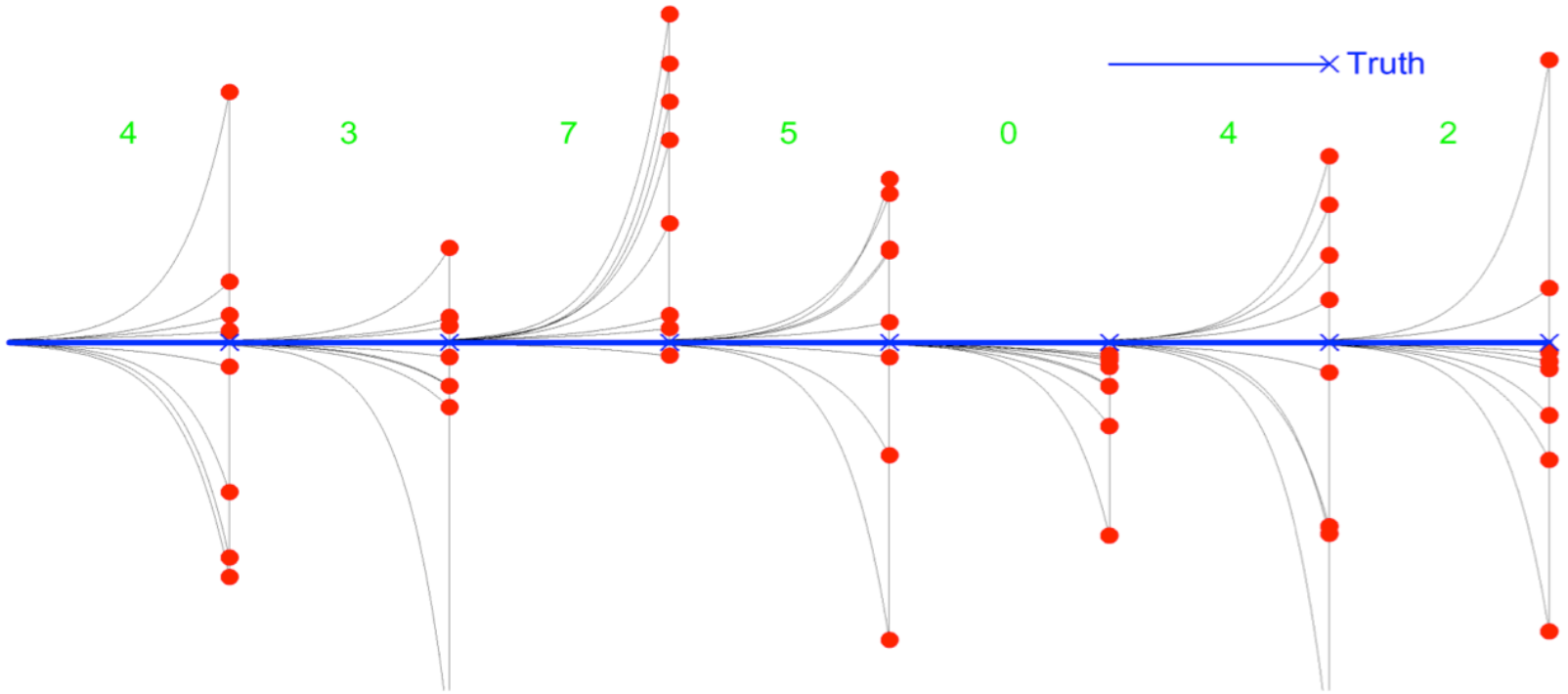

**FIGURE 2.5. A schematic of ensemble evaluation one dimension: count $N_{over}$, the number of forecasts greater than truth for each lead time. If perfect ensembles are used, then $N_{over}$ should be uniformly distributed; in $N_{exp}$ experiments, we expect the relative frequency of a particular value 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.**

# Rank Histogram




"Studies based on this approach have suggested that MMEs (CMIP3/5) are 'reliable' in that they are not too narrow or too dispersive as a sample of possible models, but existing single-model-based ensembles tend to be too narrow (Yokohata et al., 2012, 2013)."

# Rank Histogram




Rank histogram only applies to 1-D, for multi-dimensional forecast, use minimum spanning tree.

**L.A. Smith and J.A. Hensen (2004), Extending the limits of forecast verification with the minimum spanning tree.**

# Diagnostic tools

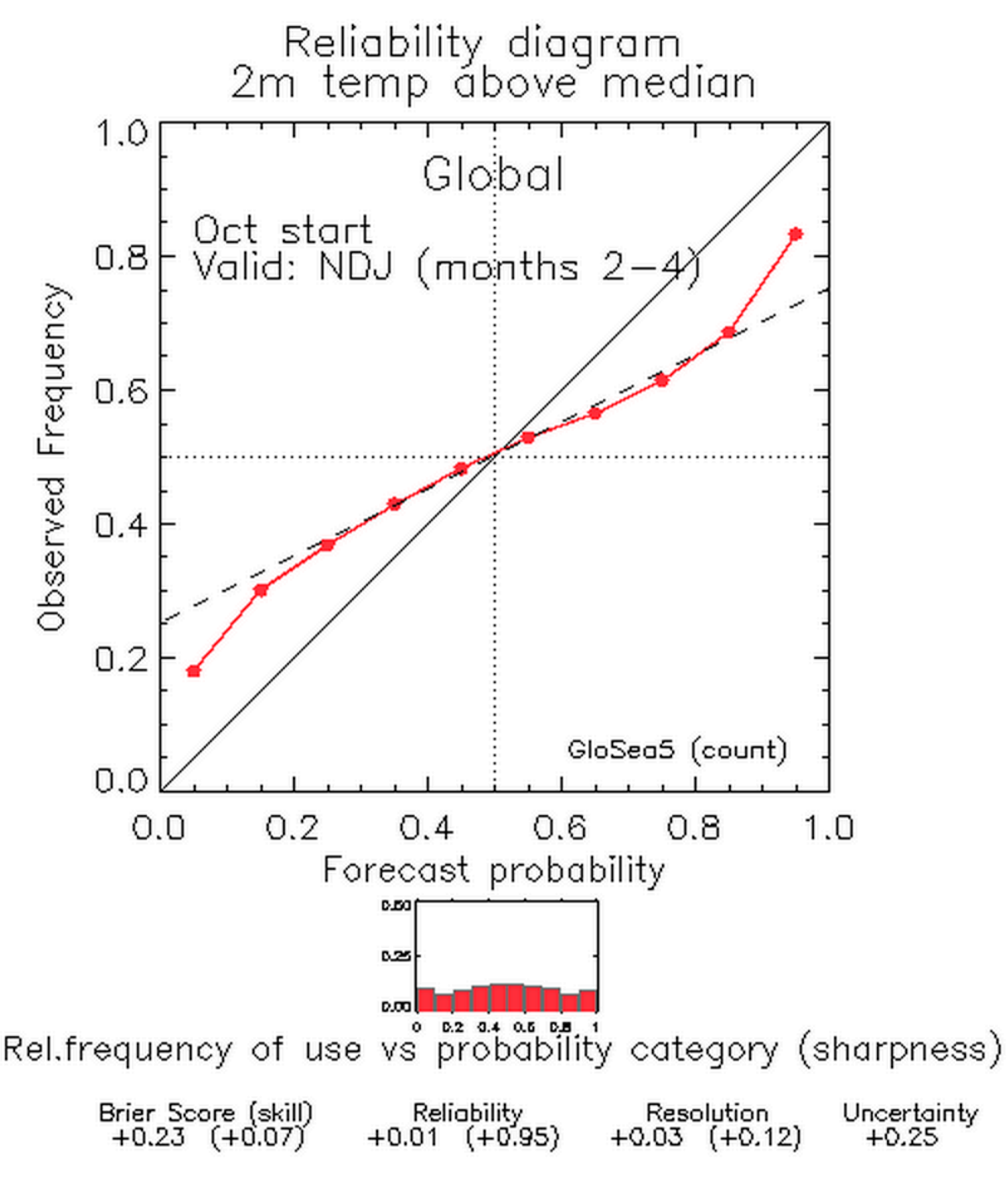
- Reliability Diagram (Calibration Graph)

For a reliable forecast one expects to see events forecast with probability *p* occur the same percentage of the time.

The calibration graph, or reliability diagram, plots the observed frequency of verifications against their forecast probability. Ideally the resulting points lie on the diagonal line.

**Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. Applied Statistics, 26(1):41–47, 1977.**

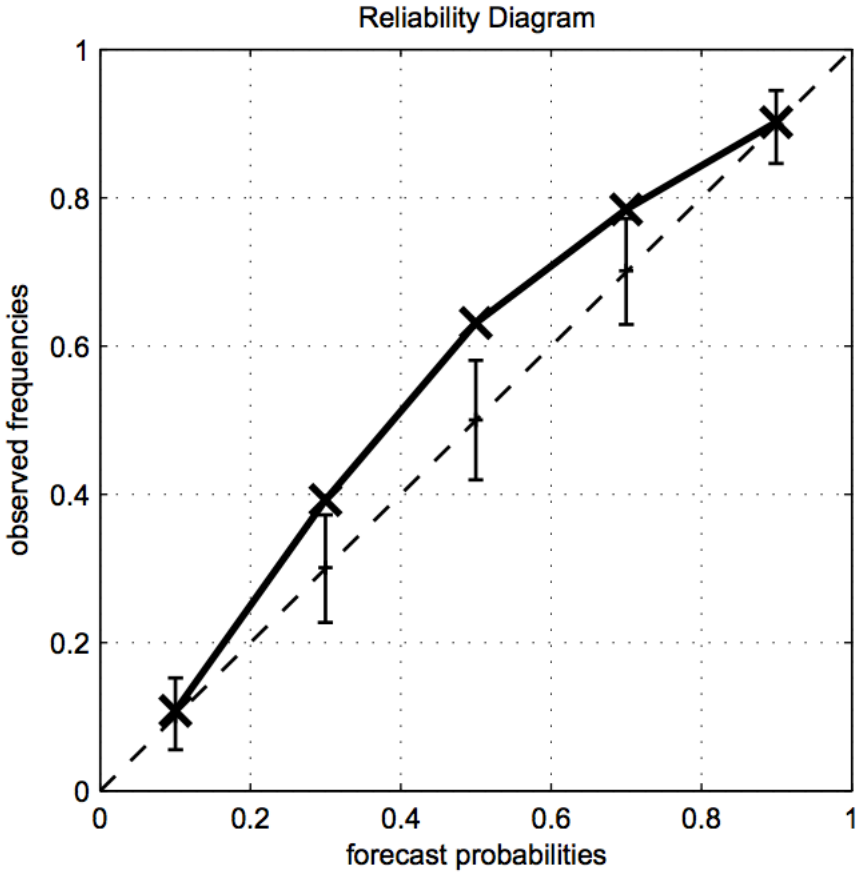
Only works on binary case




http://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/glob-seas-prob-skill

# Reliability Diagram

Consistency test is required




Jochen Bröcker, Leonard A. Smith Increasing the Reliability of Reliability Diagrams Weather and Forecasting, Vol. 22, No. 3, 2007.

# Diagnostic tools

From IPCC report Page 755:

Annan and Hargreaves (2010) have proposed a 'rank histogram' approach to evaluate model ensembles as a whole, rather than individual models, by diagnosing whether observations can be considered statistically indistinguishable from a model ensemble. Studies based on this approach have suggested that MMEs (CMIP3/5) are 'reliable' in that they are not too narrow or too dispersive as a sample of possible models, but existing single-model-based ensembles tend to be too narrow (Yokohata et al., 2012, 2013).

Rank histogram and reliability diagram are based on hypothesis test. The aim of any hypothesis test is to "reject"!


A "better" Rank histogram or reliability diagram does NOT necessarily indicate a better score.

# Skill Score

What Is A Score

A measure of "goodness"

$$S(p(x), X),$$

*p(x) = Probabilistic Forecast*

*X = Verification* 观测值

Skill of a forecast system is assessed using an *archive* of forecasts and verifications

Convention: A smaller S is better!
(the less Stupidity the better)


Who should be referee?

A qualified referee has a license.

# Strictly Proper Scores

Mathematically, a skill score is proper if for any two densities p(x) and q(x)

$$\int S(p(x), z)q(z)\mathrm{d}z \geq \int S(q(x), z)q(z)\mathrm{d}z.$$

In other words, the minimum of the left hand side over p(x) is obtained for p = q. A skill score is strictly proper if that happens only if p = q.

Essentially, this ensures that, under a proper skill score, one would only issue a forecast that one believes is right. Believing this is our best forecast should maximise our skill (by perversely minimising the expected skill score). A score being Proper prevents us from manipulating a forecast probability in order to score higher under that skill score.

Only the "true" distribution can score the highest under strictly proper scores!


# A qualified referee has a license.

The license is simply that the highest score will only be given to the “Real” Champion.

# p-score should be avoided:
# The (Naïve) Linear Score

$$S(p(x), X) = -p(X)$$

Rationale: The forecast $p$ should be large at the verification $X$

Improper

Consider an “unfair” coin, Head: 60% chance; Tail: 40% chance.

# Mean Square Error

$$S(p(x), X) = \int (X - z)^2 p(z) dz$$

*This score can be written as*

$$S(p(x), X) = (X - m)^2 + s^2$$

*Only the first and second moment of p(x) is considered*

Rationale: The mean *m* should be close to *X and the standard deviation s should be small.*

## Improper

A density p(x) centred around the mean of the true distribution and having small standard deviation would achieve a better score than the true distribution.

# Error in the mean

$$S(p(x), X) = (X - m)^2$$

*Only the first moment of p(x) is considered*

Rationale: The mean *p* should be close to *X*.

Proper, not strictly proper

In parameter estimation, using a score that is not strictly proper score may fail to locate the correct parameter value.

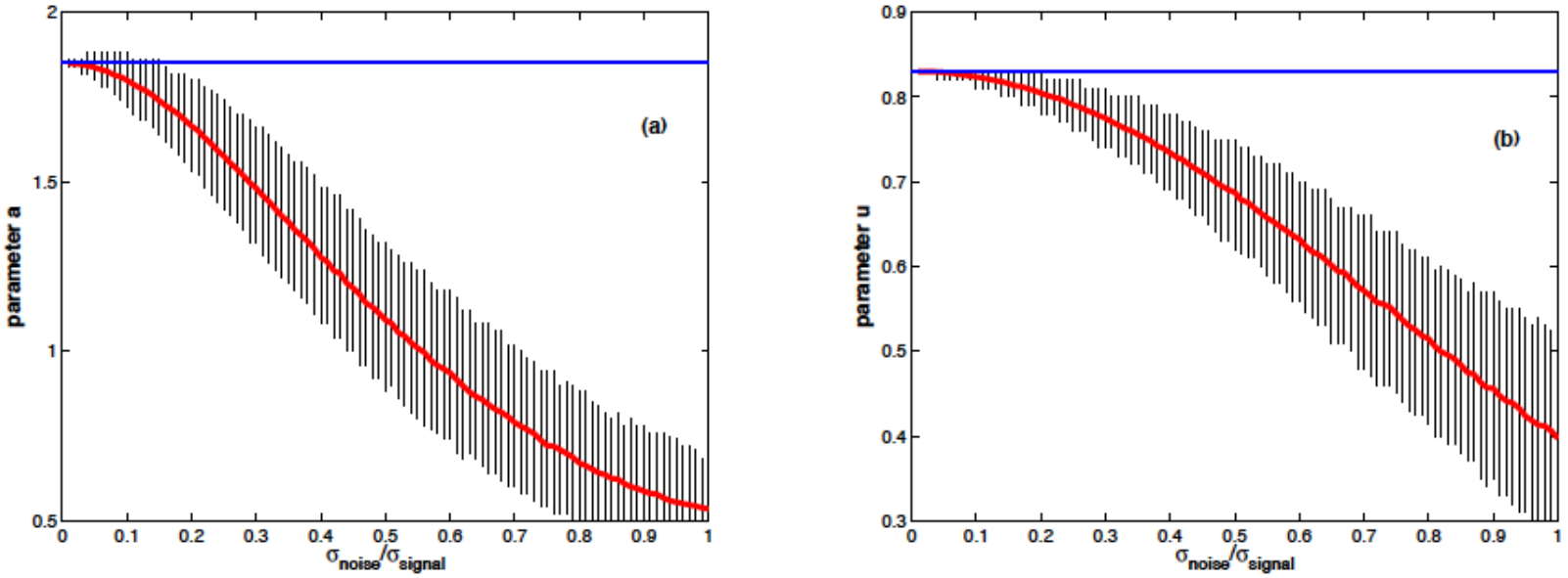



Figure 4.1: Parameter estimation using LS cost functions for different noise level, the black shading reflects the 95% limits and the red solid line is the mean, they are calculated from 1000 realizations and each cost function is calculated based on the observations with length 100, the blue flat line indicates the true parameter value (a) Logistic Map for $a = 1.85$ (b) Ikeda Map for $u = 0.83$

# The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept

By RENATE HAGEDORN*, FRANCISCO J. DOBLAS-REYES and T. N. PALMER, *ECMWF, Shinfield Park, Reading RG2 9AX, UK*

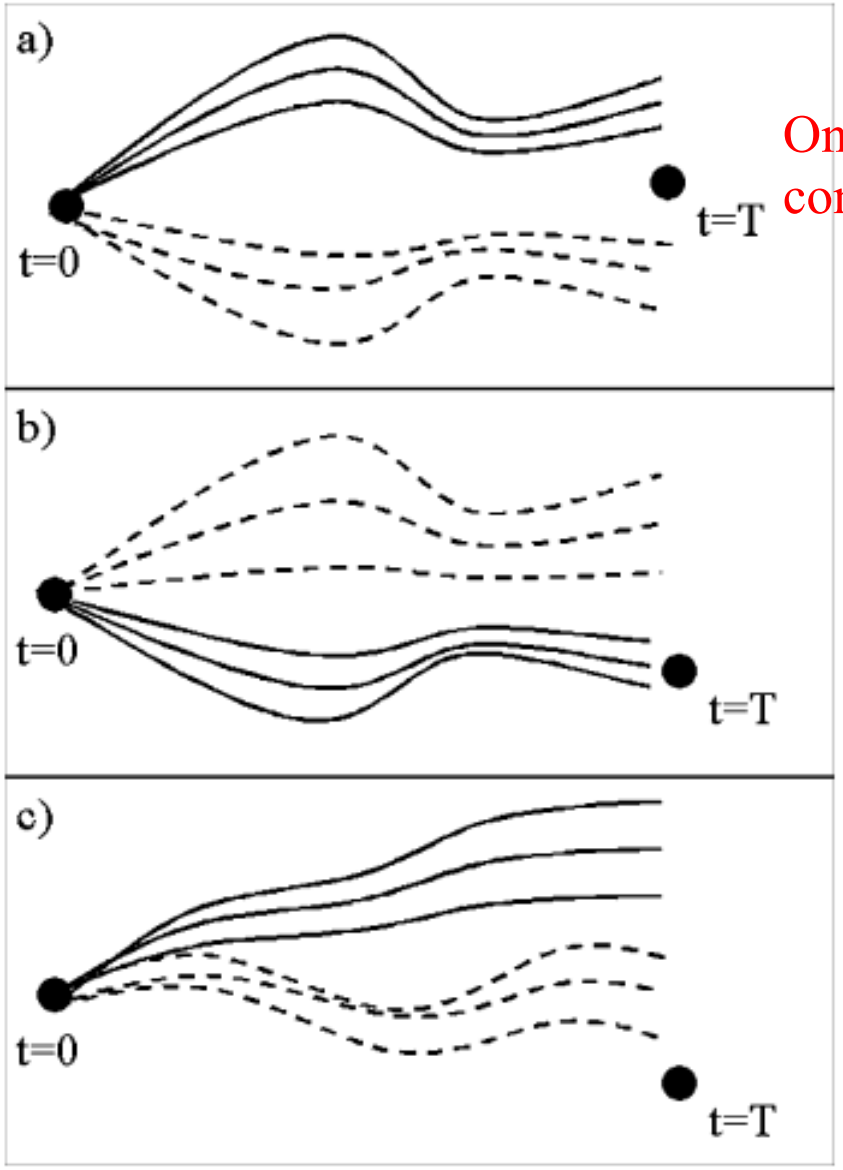



*Fig 8.* Idealized visualization of basic multi-model scenarios. For the sake of simplicity, only two single models and three ensemble members are included: model 1 (solid lines) and model 2 (dashed lines). (a) The multi-model provides the best prediction; (b) a single model provides the best prediction; (c) the verification lies outside the model predictions.

Only the ensemble mean is considered in the Mean Error.

Misguidance

# The Continuous Ranked Probability Score

The CRPS is

$$S(p(x), X) = \int [F(x) - F_0(x)]^2 dx$$

$$where\ F_0(x) = 0\ for\ x < X;\ 1\ for\ x > X$$

Strictly Proper

Rationale: The squared difference between the forecast cdf *F(x)* and the True cdf *G(x)* should be small.

# The Continuous Ranked Probability Score

The CRPS is

$$S(p(x), X) = \int [F(x) - F_0(x)]^2 dx$$

$$where\ F_0(x) = 0\ for\ x < X;\ 1\ for\ x > X$$

A generalization of the Brier score

Special case of Energy Score Family

# The Continuous Ranked Probability Score

The CRPS is

$$S(p(x), X) = \int [F(x) - F_0(x)]^2 dx$$

$$where\ F_0(x) = 0\ for\ x < X;\ 1\ for\ x > X$$

Special case of Energy Score Family

$$S_{ES}(p(x), X) = E_p \|x - X\|^\beta - \frac{1}{2} E_p \|x - x'\|^\beta$$

T. Gneiting, A. E. Raftery, (2007), Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association 102, 477:359-378

# The Proper Linear Score

The Proper Linear Score works for continuous variables:

$$X = real\ number$$

It is given by

$$S(p(x), X) = \int p(z)^2 dz - 2p(X)$$

Strictly Proper

Rationale: The squared difference between $p(x)$ and the true distribution $Q(x)$ should be small.

Special case of Power Score family

# The Proper Linear Score

The Proper Linear Score works for continuous variables:

$$X = real\ number$$

It is given by

$$S(p(x), X) = \int p(z)^2 dz - 2p(X)$$

Strictly Proper

Rationale: The squared difference between *p(x)* and the true distribution *Q(x)* should be small.

Special case of Power Score family

$$S_{PS}(p(x), X) = -\alpha p(X)^{\alpha-1} + (\alpha - 1) \int p^{\alpha}(x) dx$$

R. Selten, 1998

# The Spherical Score

The Spherical Score works for continuous variables:

$$X = real\ number$$

It is given by

$$S(p(x), X) = -\frac{p(X)}{(\int p^2(z)dz)^{1/2}}$$

Strictly Proper

Rationale: The forecast $p$ should be large at the verification $X$.

Special case of Pseudo-spherical score family

$$S_{PSS}(p(x), X) = -\frac{p(X)^{\beta-1}}{(\int p^{\beta}(z)dz)^{1/\beta}} \qquad \beta > 1$$

I.J. Good, 1971

# The Ignorance Score

The Ignorance Score works for continuous variables:

$$X = real\ number$$

It is given by

$$S(p(x), X) = -\log p(X)$$

Kullback-Leibler Inequality →Strictly Proper

Rationale: The forecast *p* should be large at the verification *X*.


A qualified referee has a license.

The license is simply that the highest score will only be given to the “Real” Champion.

What if the “Real” Champion is not in the game?

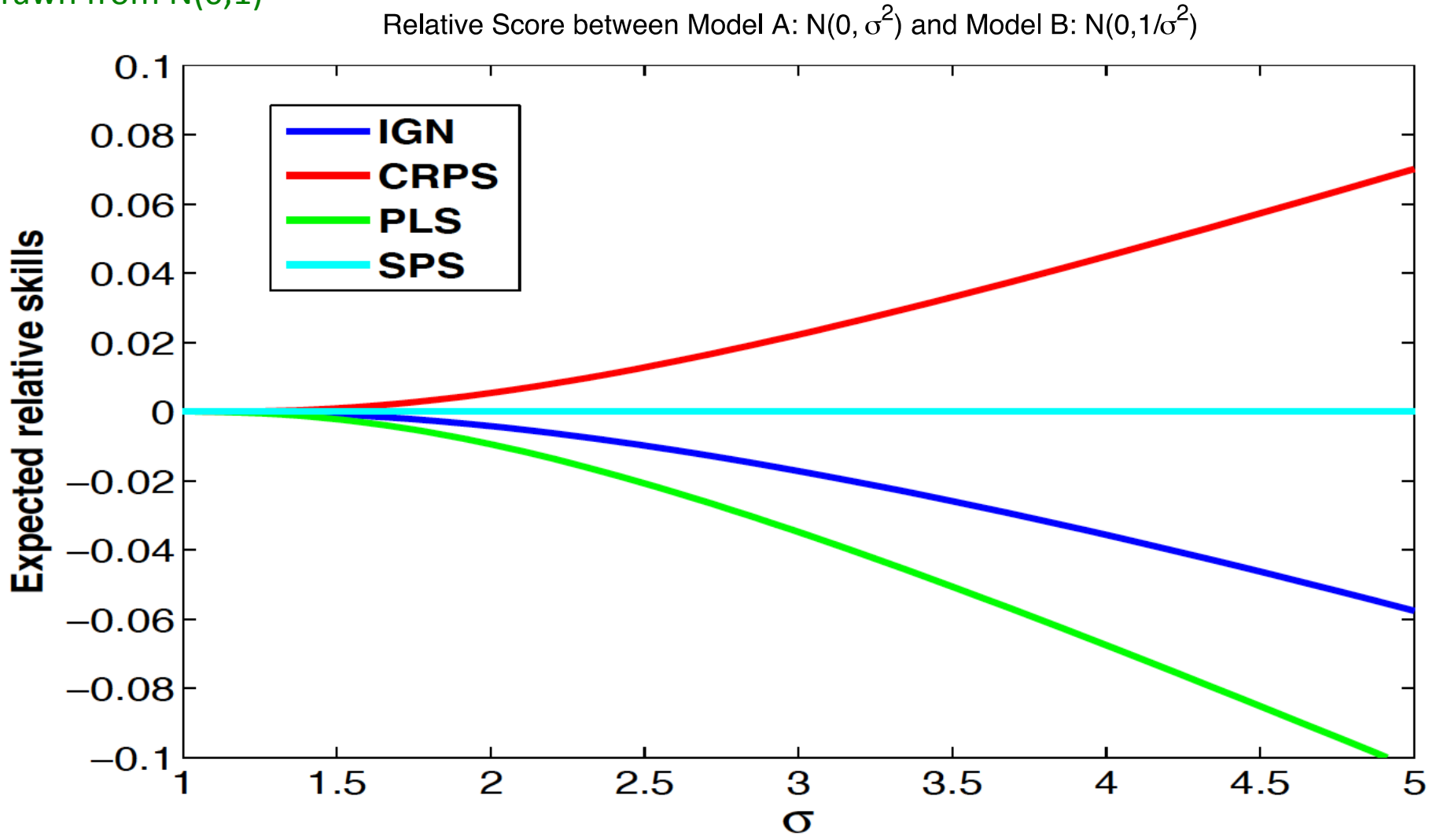
# Beyond properness

Given a (strictly) proper score, the distribution from "True" forecast system $Q$ will always be preferred whenever it is included amongst those under consideration.

When this is not the case, then even (strictly) proper scores may rank two forecast systems differently, making it difficult to provide definitive statements about forecast quality.

# Different proper skill scores might rank models different

Verifications are drawn from N(0,1)



# Beyond properness

Given a strictly proper score, the distribution from "True" forecast system Q will always be preferred whenever it is included amongst those under consideration.

When this is not the case, then even strictly proper scores may rank two forecast systems differently, making it difficult to provide definitive statements about forecast quality.

There are infinite number of strictly proper skill scores.

Power Scores; Energy Scores…

Which score shall we choose?

# Locality

$$Ignorance = -log(p(X))$$

Local

Here only $p$(verification) matters

$$S(p(x), X) = \int p(z)^2 dz - 2p(X)$$

Non-local

Here the whole shape of $p$ matters! Thus the score depends on "events" that never happen.

# Locality

A local skill score gives no credit for probability density which is "close to" the verification, X.

A local score depends only on p(X).

Ignorance is the only proper, local skill score for continuous variables.

There are infinite number of non-local strictly proper scores

# Score interpretation: how much better?

- Proper Linear Score

$$S(p(x), X) = \int p(z)^2 dz - 2p(X)$$

$$E(S_{PL}(p(x), X)) = \int [Q(X) - p(X)]^2 dX - \int Q^2(X) dX$$

- Continuous Ranked Probability Score

$$S(p(x), X) = \int [F(x) - F_0(x)]^2 dx$$

$$E(S_{CRPS}(p(x), X)) = \int [F(z) - G(z)]^2 dz + \int G(z)[1 - G(z)] dz$$

Interpretations of nonlocal scores are tied up with the true forecast system, which is in fact unknown.

**Broecker, J. (2009) Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society, 135 (643).**

# Interpretation of Ignorance

$$S(p(x), X) = -\log p(X)$$

- (Relative) Ignorance can be interpreted as how much more/less probability mass Model A assigns to the verification than Model B.
- Ignorance can be related to the amount of information expected from a forecast. (Shannon's information entropy)
- Ignorance describes the expected rate at which the forecaster's wealth changes with time, under Kelly betting scenario.
- Ignorance can be easily communicated an effective interest rate.

# A score is *Implausible:*

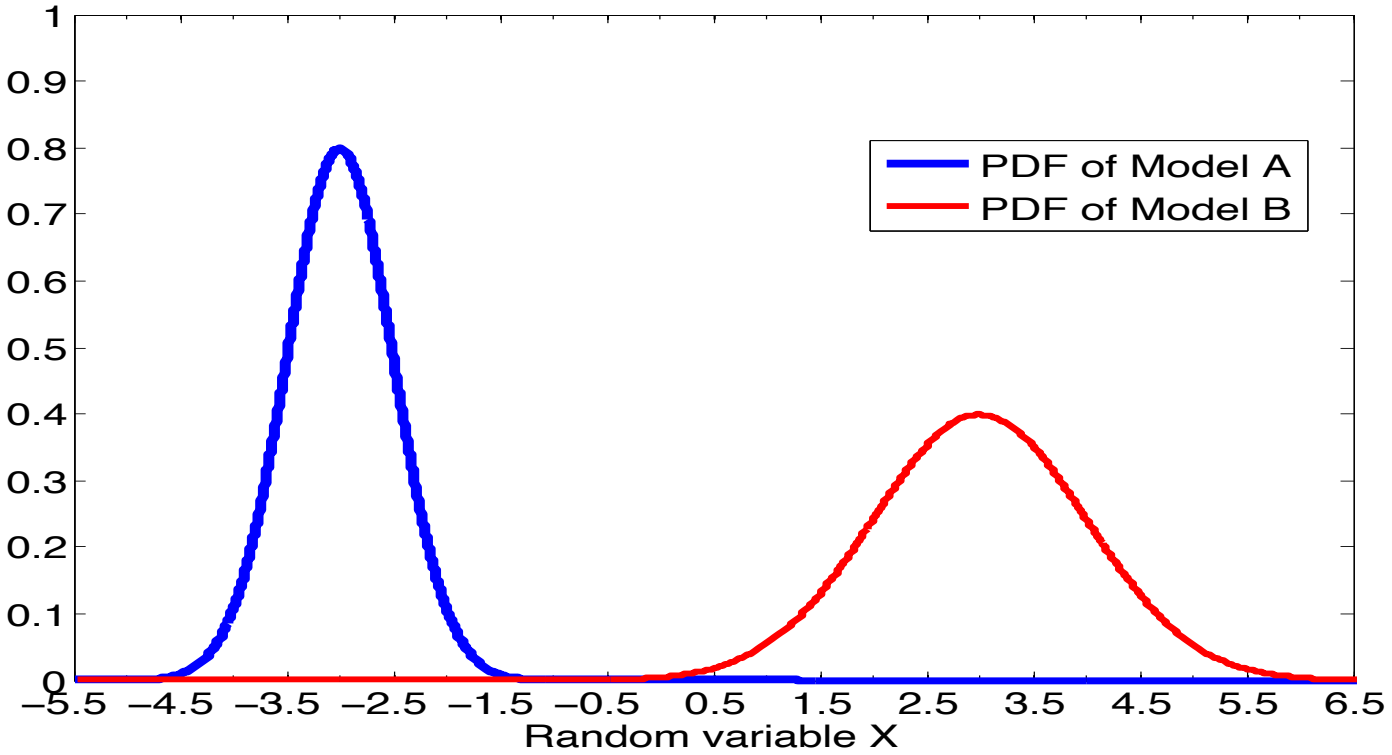
if for ANY $r > 1, r \in \mathbb{R}$

there exist two forecast systems $p(x)$ and $q(x)$, and $Y$, where

$$p(Y)/q(Y) = r \text{ while } S(p, Y) > S(q, Y)$$

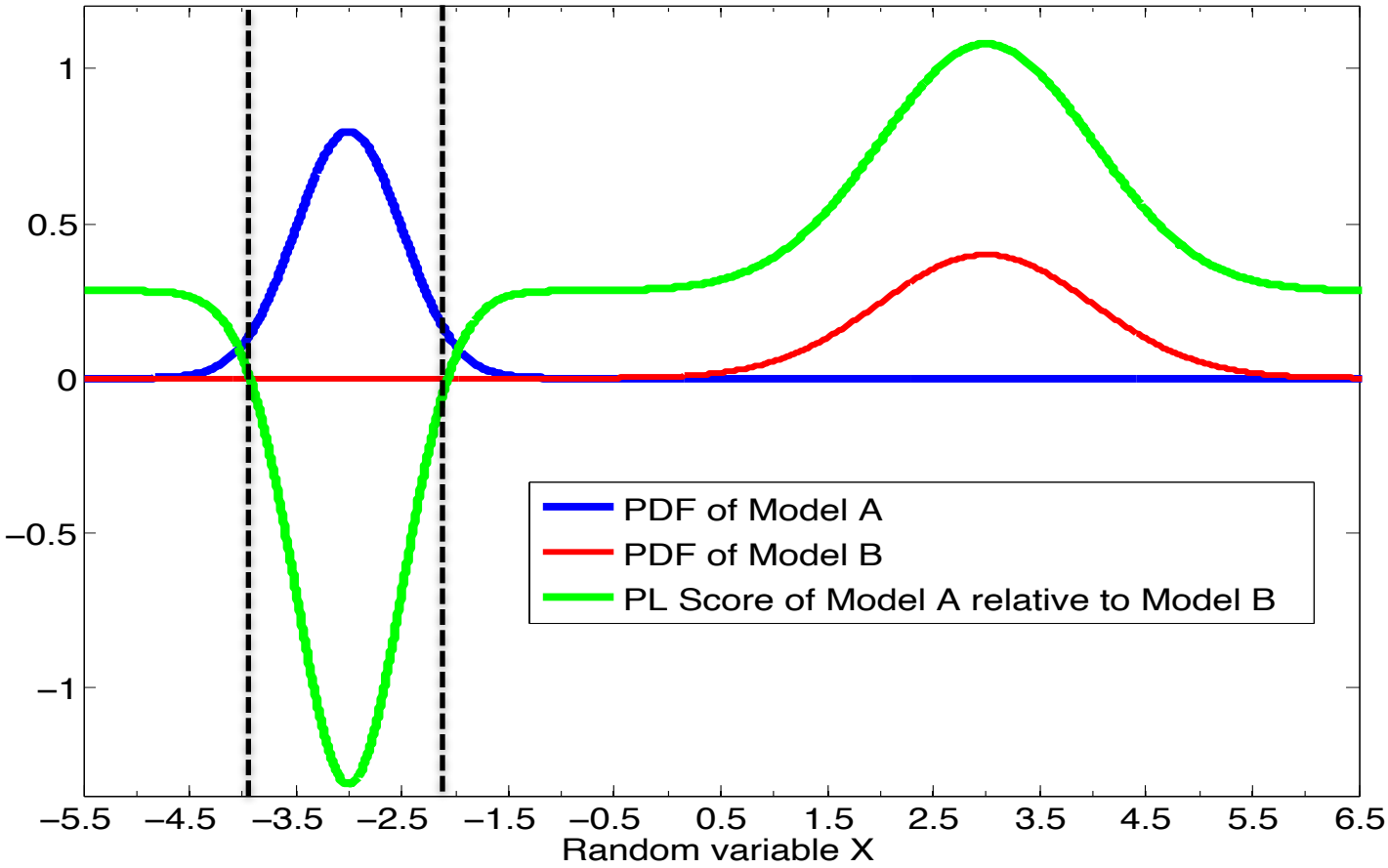
Du, H. (2021). Beyond Strictly Proper Scoring Rules: The Importance of Being Local, Weather and Forecasting, 36(2), 457-468. Retrieved Aug 23, 2021

# Non-Local scores are “implausible”



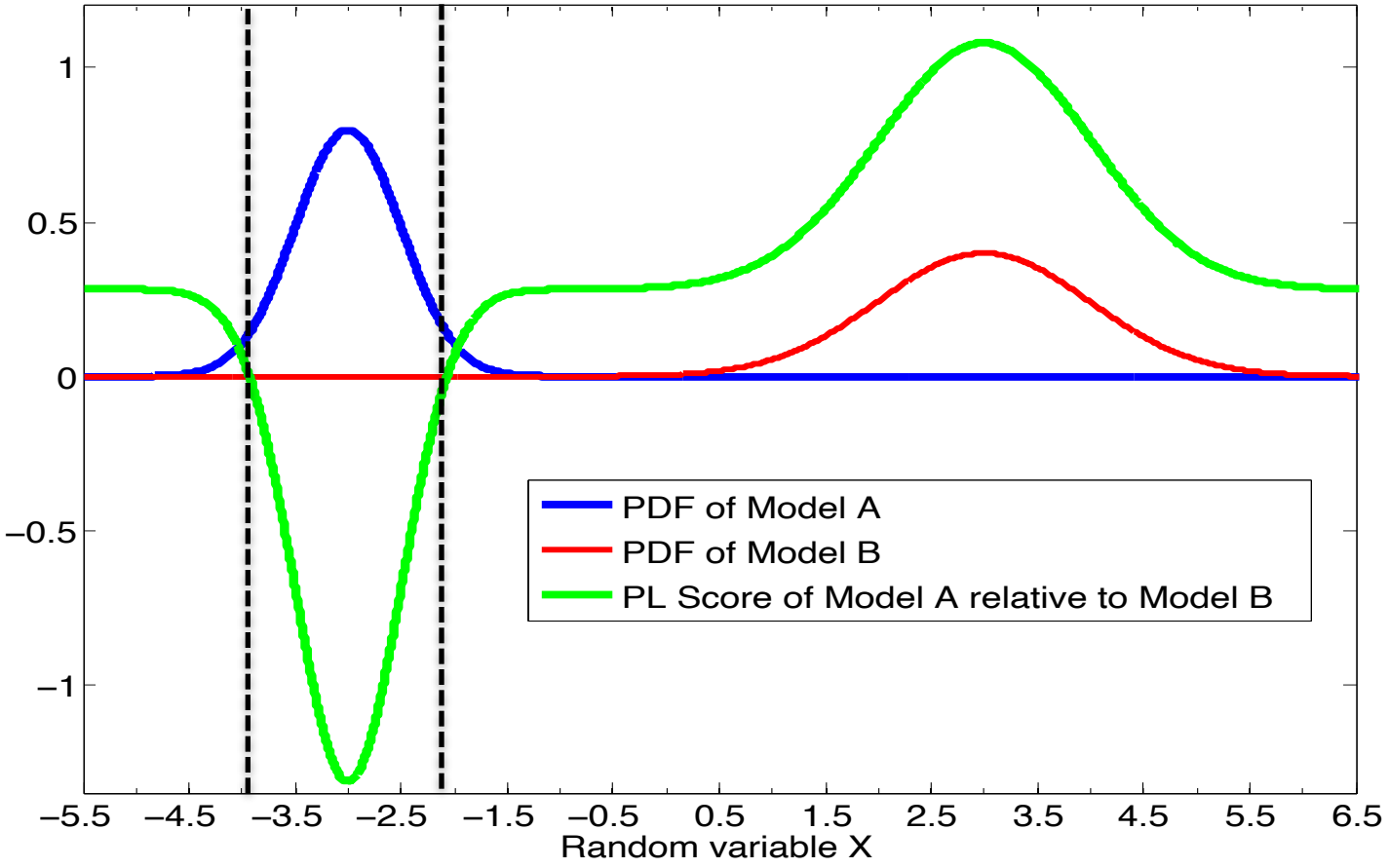

If Model A is better than Model B, where would one expect the verification land (consider the “true” pdf is simply a delta function)?

# Proper Linear score is “implausible”



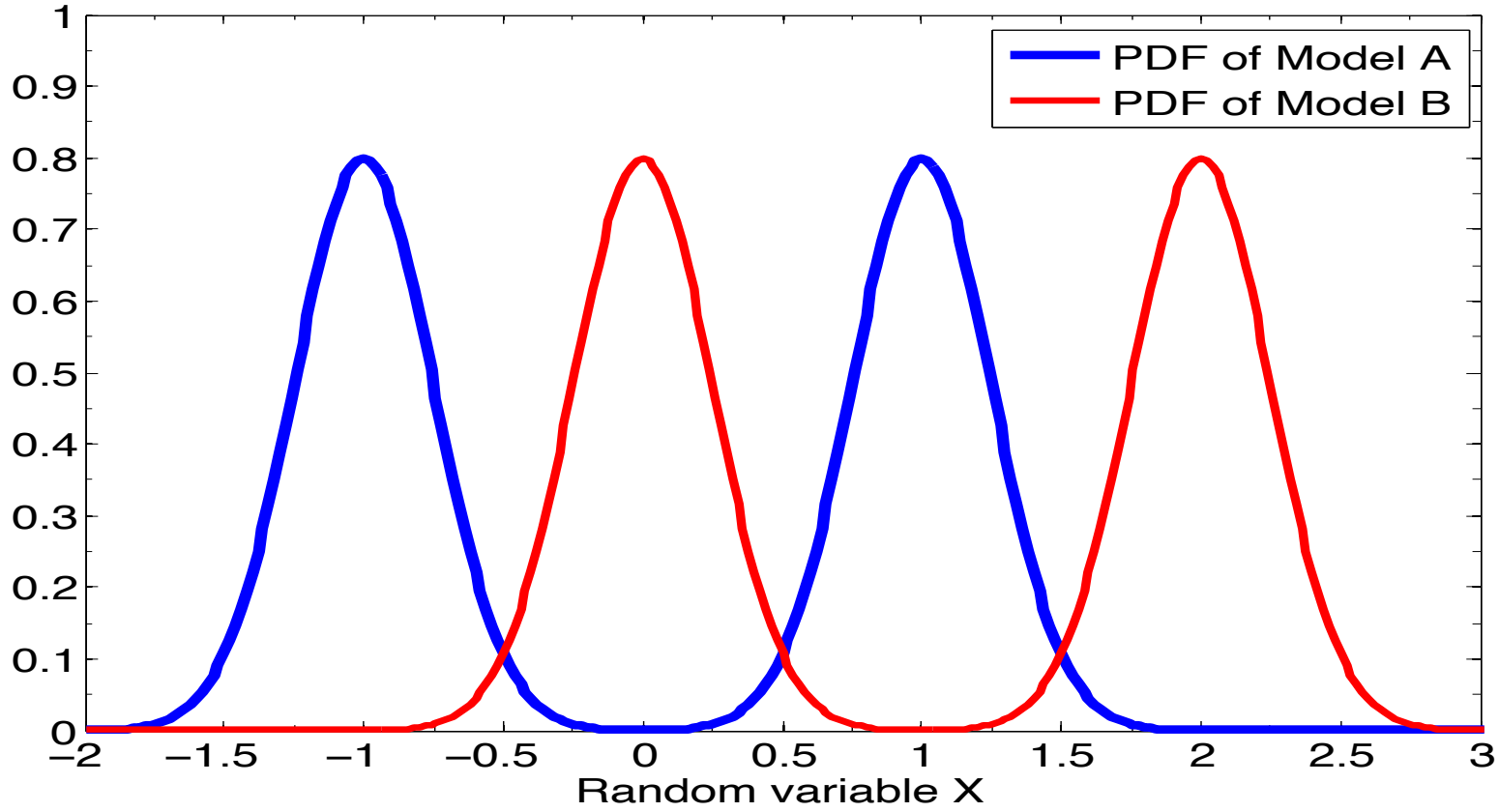

If Model A is better than Model B, where would one expect the verification land (consider the “true” pdf is simply a delta function)?

# Proper Linear score is “implausible”



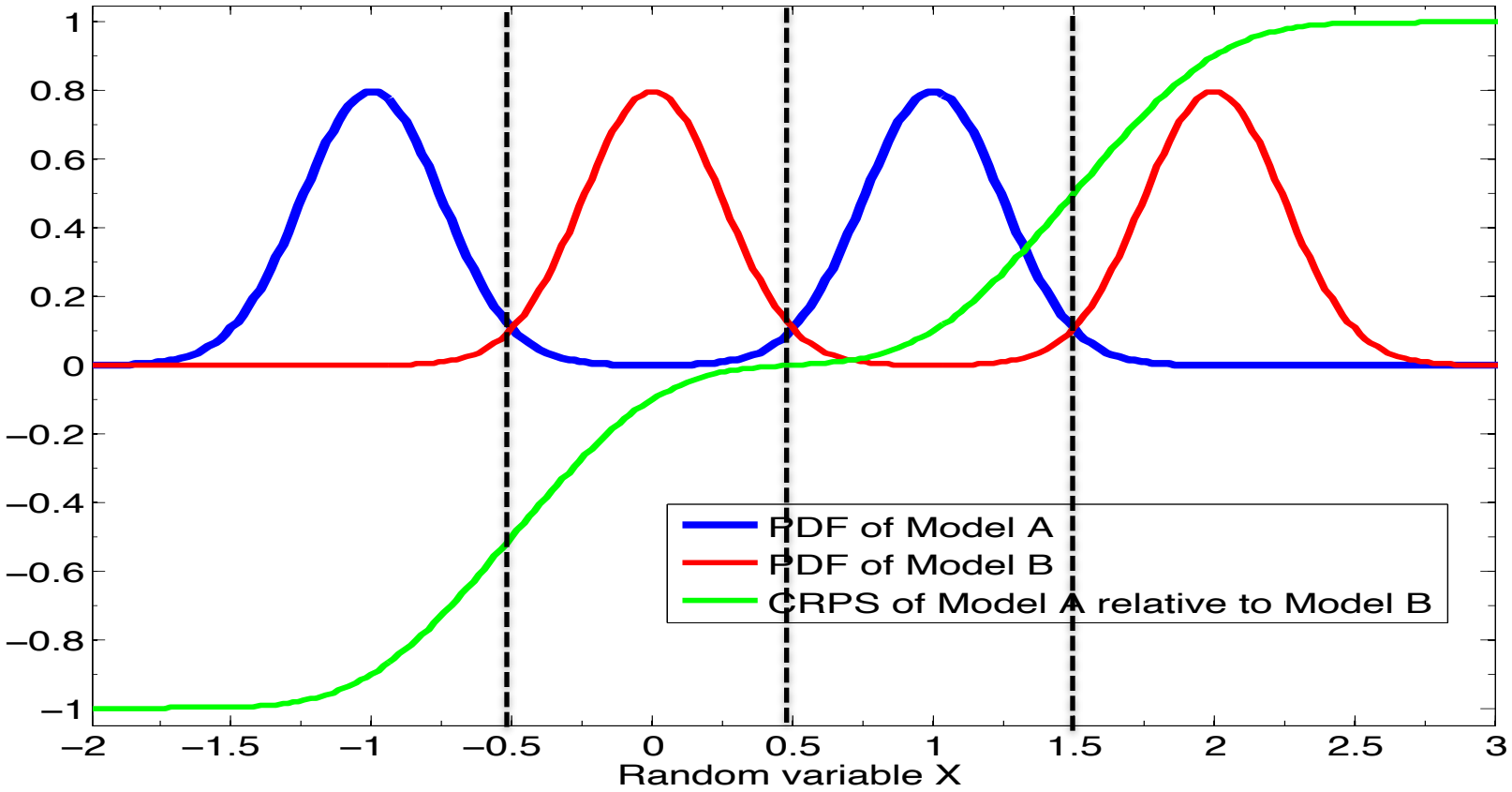

$$S(p(x), X) = \int p(z)^2 dz - 2p(X)$$

# Continuous Rank Probability score is “implausible”



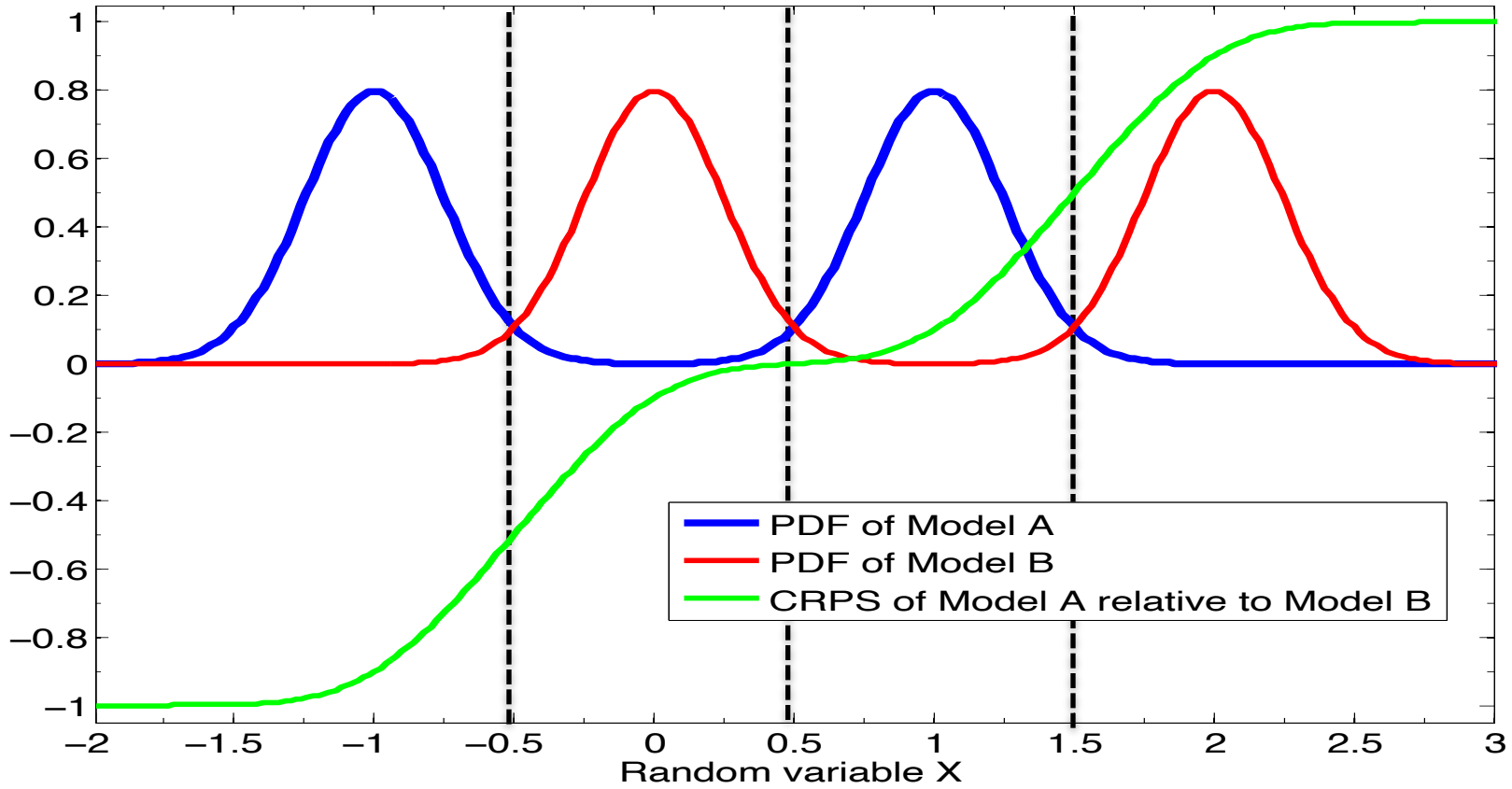

If Model A is better than Model B, where would one expect the verification land (consider the “true” pdf is simply a delta function)?

# Continuous Rank Probability score is “implausible”




If Model A is better than Model B, where would one expect the verification land (consider the “true” pdf is simply a delta function)?

# Continuous Rank Probability score is “implausible”




$$S(p(x), X) = \int [F(x) - F_0(x)]^2 dx$$

The optimal CRPS is achieved when *X* is the median of the forecast distribution *p(x)*

# Scoring Rules Under Transformation

In practice, it is common that the variable of interest is not the variable observed but a function of the observed variable.

wind speed V ⟶ wind power $\propto V^3$

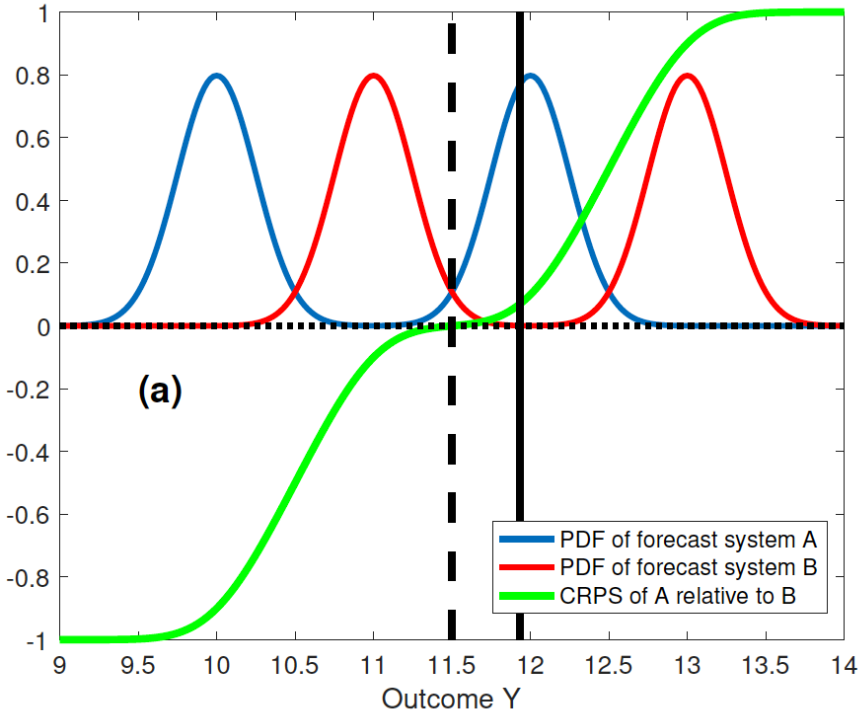
It is desirable for a scoring rule to provide coherent evaluations before and after a smooth transformation being applied to the forecast variable.

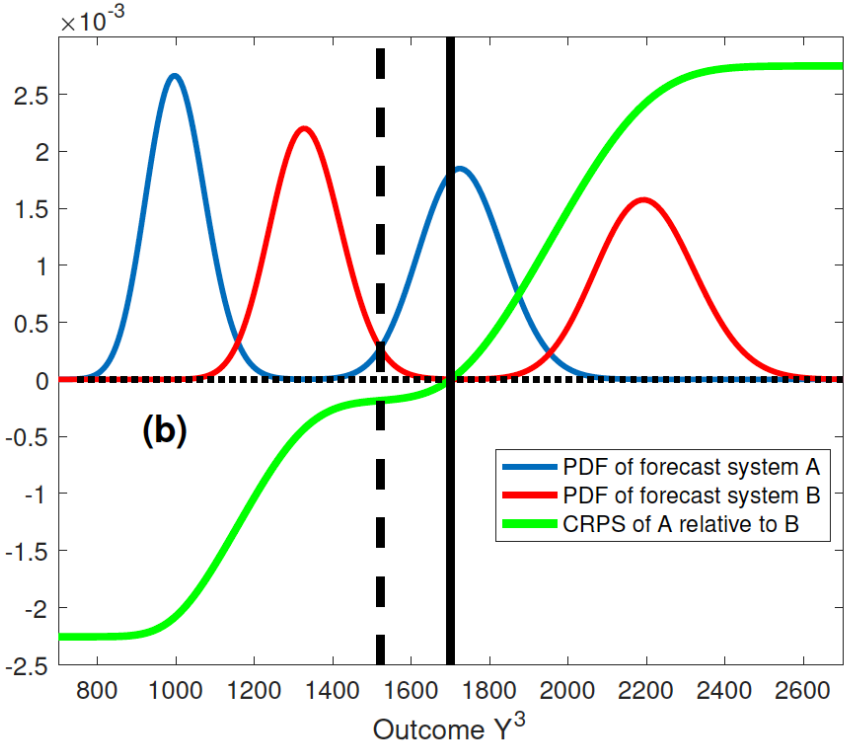
Ignorance score is invariant under smooth transformation.

Nonlocal scores are NOT…..

# Scoring Rules Under Transformation

wind speed V ⟶ wind power $\propto V^3$




The black dashed vertical line and solid line in (a), where Y = 11.5 and Y = 11.94 respectively, corresponds to those in (b), where Y = $11.5^3$ and Y = $11.94^3$.

Du, H. (2021). Beyond Strictly Proper Scoring Rules: The Importance of Being Local, Weather and Forecasting, 36(2), 457-468. Retrieved Aug 23, 2021

# Interpretation of Ignorance

A useful interpretation of ignorance can be found in gambling. If a gambler is able to stake an arbitrary fraction of their wealth on outcome $i$ then, to maximise their expected return averaged over sequential bets, a gambler should bet proportionally. More precisely, a fraction $w_i$ of ones wealth, where $w_i$ is the forecast probability of event $i$ occurring, should be wagered on the $i^{th}$ outcome. This strategy is called Kelly Betting.

Given this strategy, the ratio of the gamblers' wealth after the bet to that before the bet has an expected value of $2^W$ where

$$\begin{aligned} W &= \sum_{i=1}^{n} p_i \log_2 o_i w_i \\ &= \sum_{i=1}^{n} p_i \log_2 w_i + \sum_{i=1}^{n} p_i \log_2 o_i, \end{aligned}$$

$o_i$ is the odds assigned to outcome $i$ and $p_i$ is the true probability of event $i$ occurring.

# Interpretation of Ignorance

Given this strategy, the ratio of the gamblers' wealth after the bet to that before the bet has an expected value of $2^W$ where

$$\begin{aligned} W &= \sum_{i=1}^{n} p_i \log_2 o_i w_i \\ &= \sum_{i=1}^{n} p_i \log_2 w_i + \sum_{i=1}^{n} p_i \log_2 o_i, \end{aligned}$$

$o_i$ is the odds assigned to outcome $i$ and $p_i$ is the true probability of event $i$ occurring.

If the house odds are based on its own forecast probability distribution $g_i$ then $o_i = 1/g_i$ and equation 4 becomes

$$W = E[IGN]_{house} - E[IGN]_{gambler},$$

where $E[IGN]$ is the expected value of the ignorance. One can only expect to make money if the model employed has a lower ignorance than that generating the odds. The difference in $IGN$ reflects the wealth doubling (or halving) time of the gambler relative to the house.

# Communicating Skill

But given that we know our models are imperfect, should we be interpreting these *predictive distribution functions* as pdfs at all?