The Strength of Ensembles Lies not in Probability Forecasting

How can one best use an ensemble forecast system in making decisions in the real world that are influenced by the future weather? Several actual applications will be considered, and some real-time forecasting will be required (interactively) form the audience. It will be argued that it is costly to act as if ensembles gave us useful probabilities (in any of the Bayesian senses), but that ensemble can and do yield probabilistic information and can and has been used to advantage in weather sensitive decision making. Ensembles can provide early warning that our model is sensitive to the state of the atmosphere today, but that is a somewhat different from any claim regarding the predictability of the atmosphere itself today. The search for accountable ensembles (Smith, 1995) is, I now believe, wrong-headed, given that our dynamical models are imperfect. Rather than assuming calibration where it rarely exists, one can work with practitioners to identify useful questions which can be informed in a robust and useful manner. The Forecast Direction Error approach illustrates one successful application in the electricity sector (Smith, 2016). Our approach can never be as attractive as what one could achieve given "true" (or accountable) probability forecasts, but then we are not competing against such "fantastic objects." Implications for other uses of ECMWF forecasts, and for model development, are touched on.

Slido.com
#D571

Smith, L.A. (1995) 'Accountability and error in ensemble forecasting', In 1995 ECMWF Seminar on Predictability. Vol. 1, 351-368. ECMWF, Reading.

Smith, L.A. (2016) 'Integrating information, misinformation and desire: improved weather-risk management for the energy sector', in Aston, P et al. (ed.) *UK Success Stories in Industrial Mathematics*, 289-296. Springer

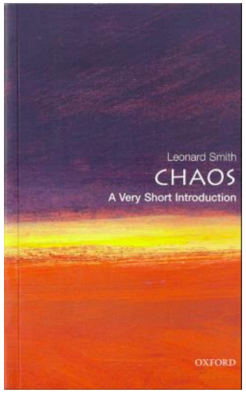
CATS CENTRE FOR THE ANALYSIS OF TIME SERIES

LSE CATS CENTRE FOR THE ANALYSIS OF TIME SERIES

www.lsecats.ac.uk

The Strength of Ensembles Lies not in Probability Forecasting:

Information for Decision Support



Leonard Smith
London School of Economics
& Pembroke College, Oxford

CRUISSE

GLIMPSE

This Talk Would Not Be THIS Talk without:

ENSEMBLES

Slido   www.slido.com #D571

If you want to ask questions (or answer mine) or just lurk and see what other people ask, then on your “mobile device” go to:

www.slido.com   Meeting #D571

Please go there now if you want too! The meeting will be open for 6 days and CATS will respond to (if not answer) each question posted.

The meeting number is also on my last slides.

Slido.com
#D571

Just Enough Decisive Information (JEDI)

The original aim of "weather forecasting" was to warn of the weather thought probable.

Then the aim was to say what the weather would be.

When this was deemed impossible in principle, the aim shifted to early warning, then accountable probability forecasts of the weather. (Back to Galton vs. Fitzroy.)

I believe that we are now at another such junction, but we do not have a well defined mathematical target.

For *users* of forecasts, I suggest we call this aim "just enough decisive information."

Information which aides decision making,
but does not make it w-trivial.

Slido.com
#D571

Probability and Ensembles

We are only interested in forecasts of empirically observable events, events in the real world.

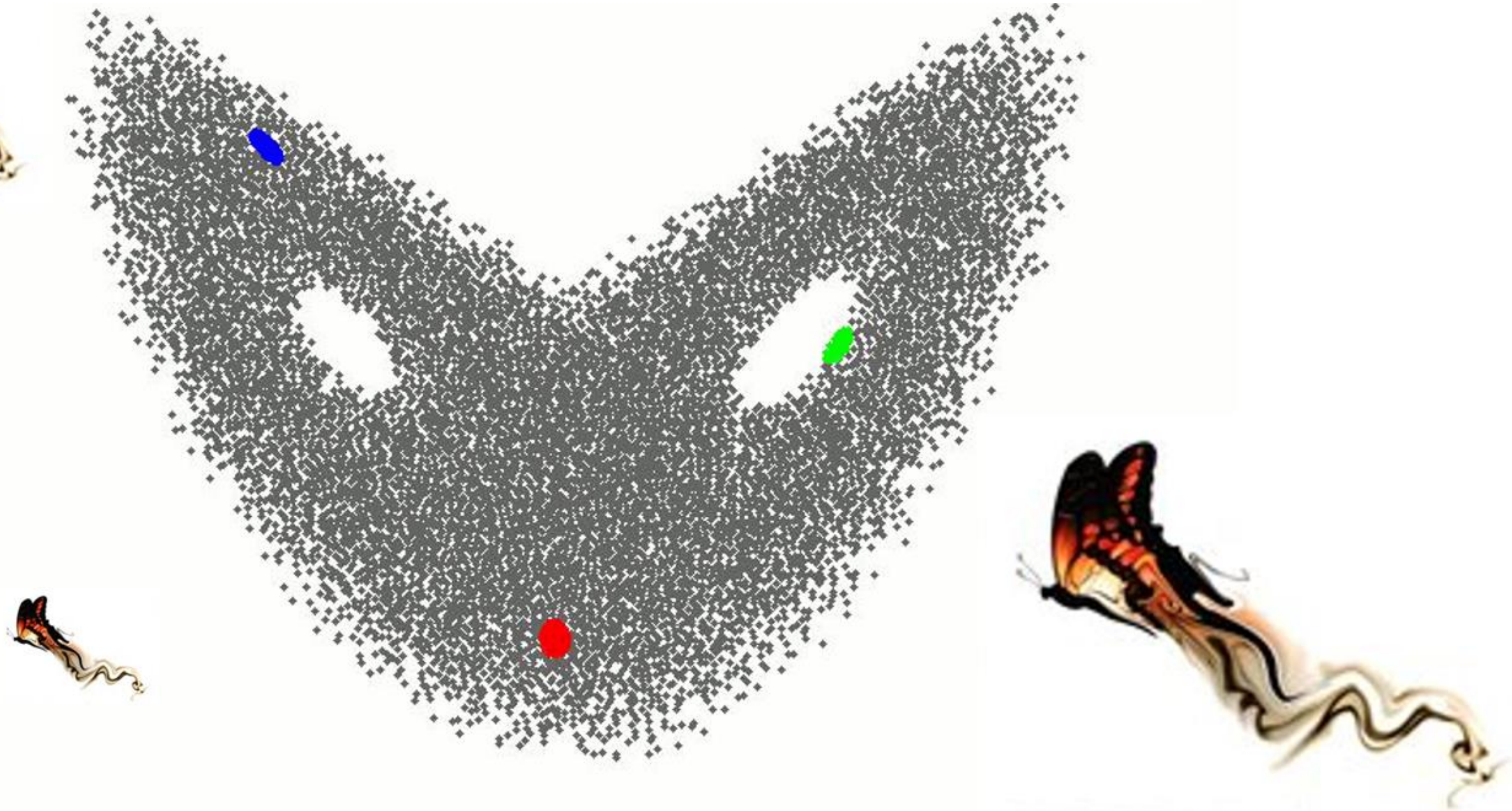
Ensembles exist in model-land. We must “interpret” ensembles to get relevant distributions in the real-world.

There are good mathematical reasons for believing we can never get accountable probability forecasts from our mathematical models.

Consider this illustration…

Predictability and Chaos

Skill Today, Gone Tomorrow

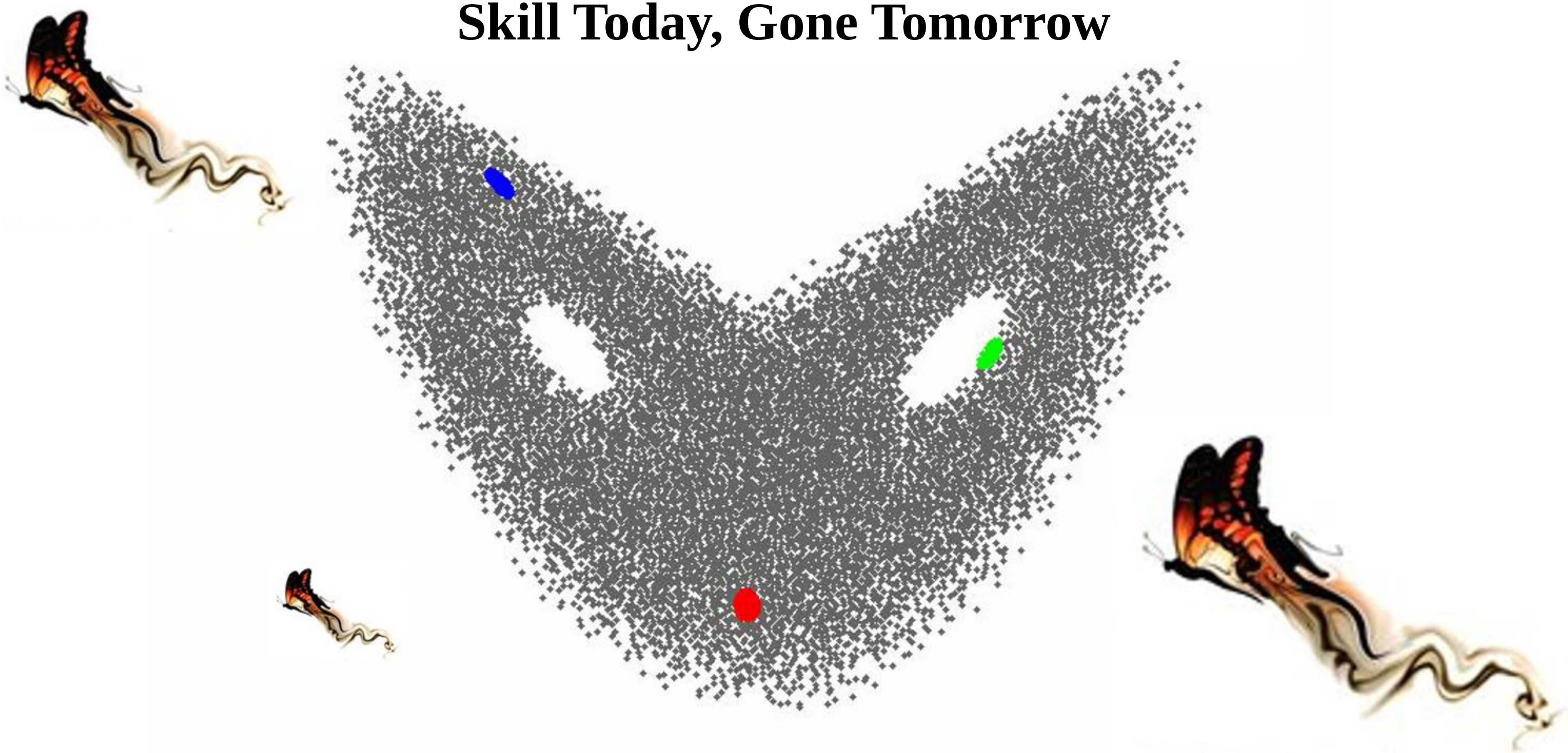


Some days we have more skill than average, some days less.

The hope is for ensembles to inform us which is which, in advance!

Predictability and Chaos

Skill Today, Gone Tomorrow

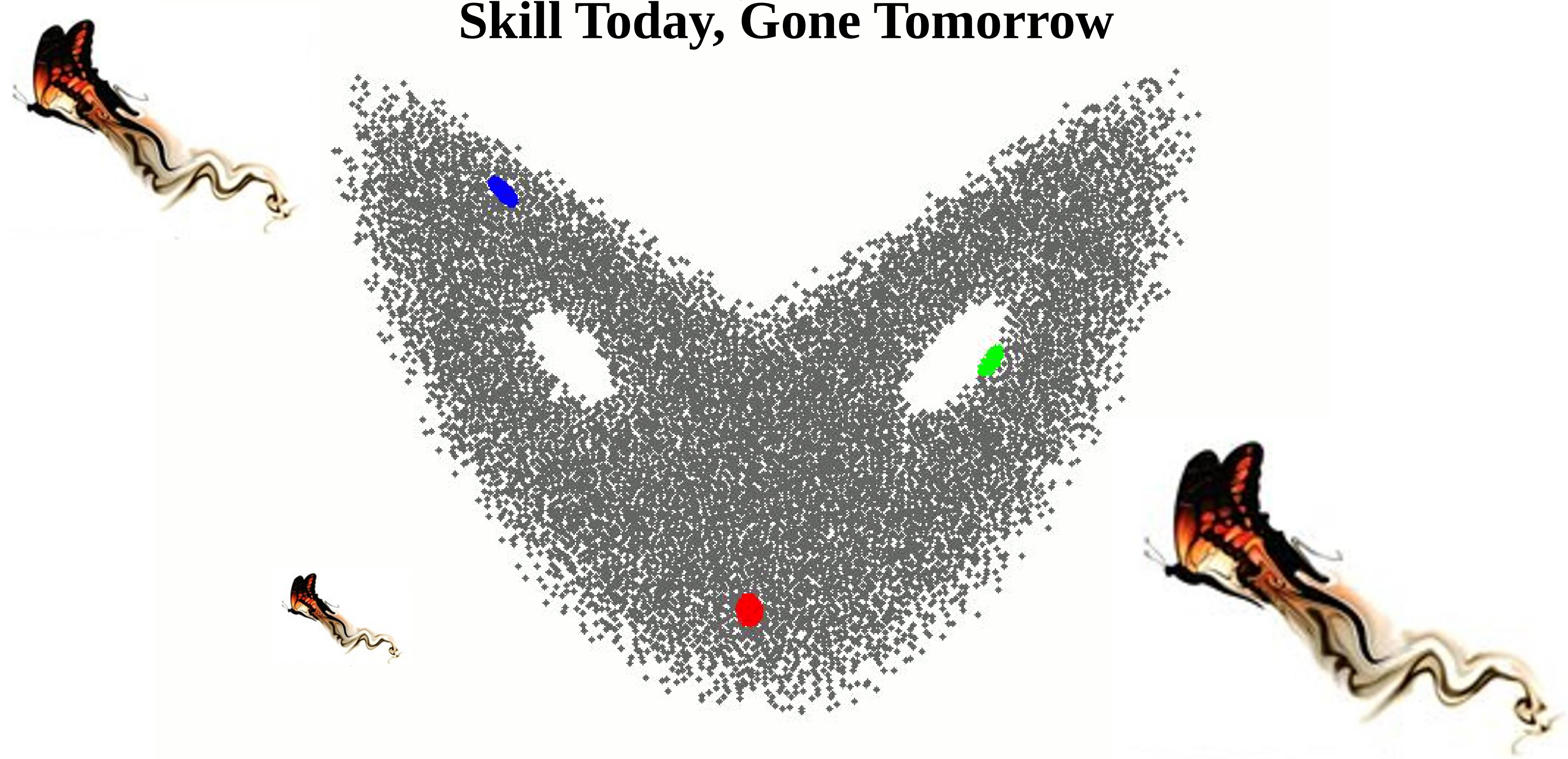


Some days we have more skill than average, some days less.

The hope is for ensembles to inform us which is which, in advance!

Predictability and Chaos

Skill Today, Gone Tomorrow

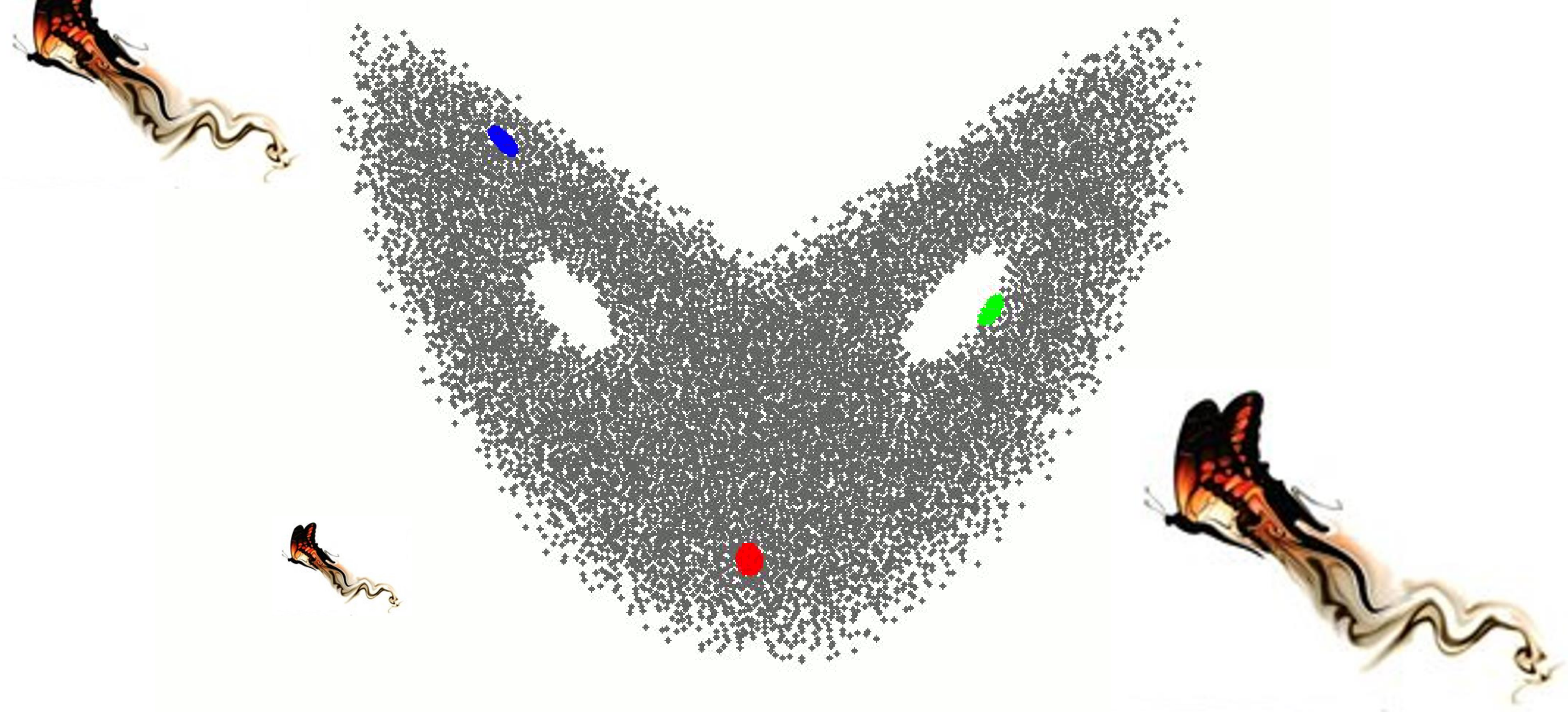


Some days we have more skill than average, some days less.

The hope is for ensembles to inform us which is which, in advance!

Predictability and Chaos

Skill Today, Gone Tomorrow



Some days we have more skill than average, some days less.

The hope is for ensembles to inform us which is which, in advance!

Kobayashi Maru

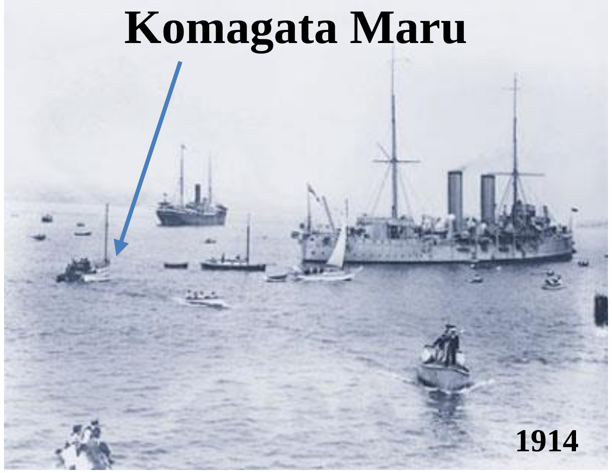
As long as you stay in model land,
you can do anything.

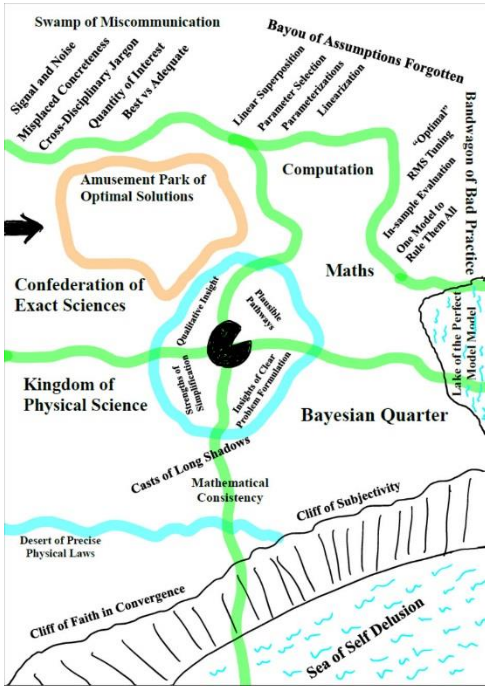
We build extremely complicated models,
to predict the weather, to drive cars, make unstable planes fly, for nuclear stewardship…

These model produce useful information regarding the real world, but are imperfect.

It is good fiction to re-write code to improve the outcome (in “fictional model-lands”).
This fails even in “fictional real-worlds.”

It is poor science, poor engineering and disastrous policy making to believe reality has rewritten itself to describe your model.




Swamp of Miscommunication
Signal and Noise
Misplaced Concreteness
Cross-Disciplinary Jargon
Quantity of Interest
Best vs Adequate
Bayou of Assumptions Forgotten
Linear Superposition
Parameter Selection
Parameterizations
Linearization
Bandwagon of Bad Practice
“Optimal”
RMS Tuning
In-sample Evaluation
One Model to Rule Them All
Amusement Park of Optimal Solutions
Computation
Maths
Confederation of Exact Sciences
Qualitative Insight
Plausible Pathways
Lake of the Perfect Model Model
Kingdom of Physical Science
Strengths of Simplification
Insights of Clear Problem Formulation
Bayesian Quarter
Casts of Long Shadows
Mathematical Consistency
Cliff of Subjectivity
Desert of Precise Physical Laws
Cliff of Faith in Convergence
Sea of Self Delusion

Fewer Model Intercomparison Projects (MIPs)
More Reality Intercomparison Projects (RIPs)

Predictability and Structural Model Error

Systems/model pairs

Model

$$\begin{aligned} \dot{x} &= -\sigma x + \sigma y \\ \dot{y} &= -xz + rx - y \\ \dot{z} &= xy - bz \end{aligned}$$

System

c sin(x/c)

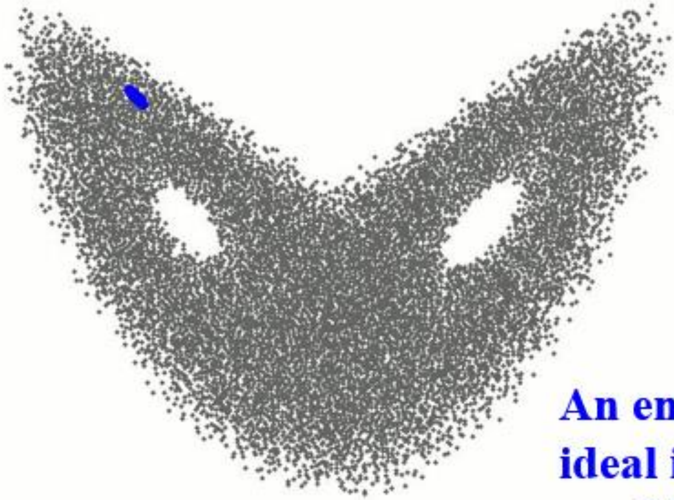
c = 128

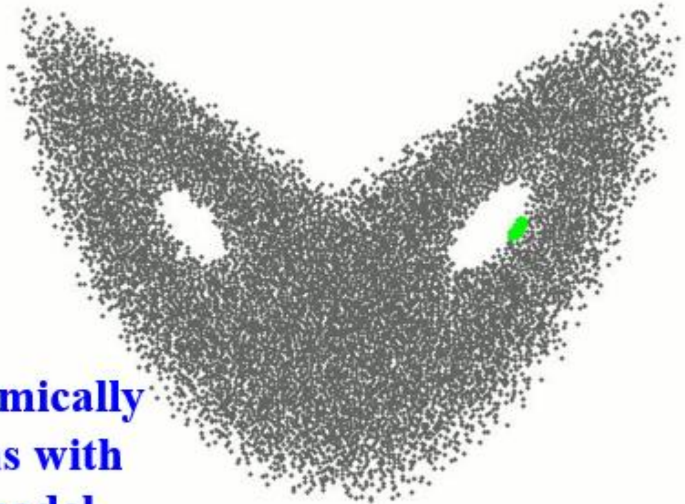
$$\begin{aligned} \dot{x} &= -\sigma x + \sigma y \\ \dot{y} &= -xz + rx - y \\ \dot{z} &= xy - bz \end{aligned}$$

This is **Structural Model Error.**

Predictability and Structural Model Error

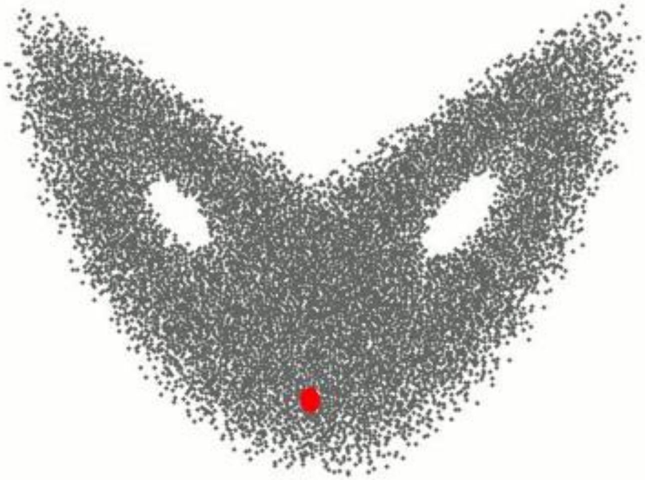
$x \rightarrow c\ \sin(x/c)$ on RHS with $c=128$





An ensemble of dynamically ideal initial conditions with good but imperfect model

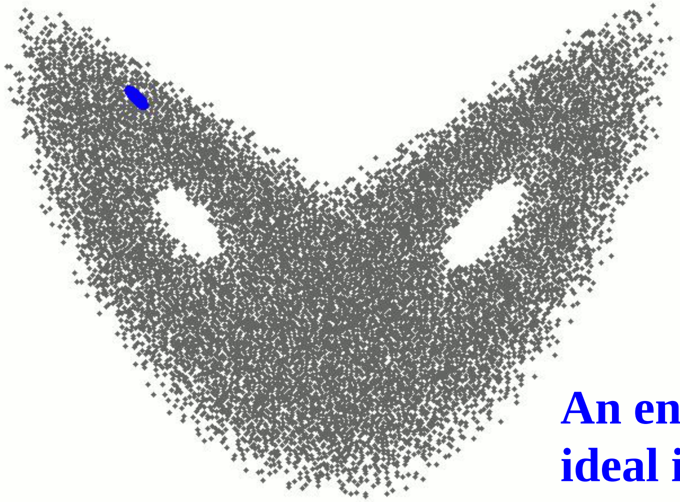
Model may shadow the system for an arbitrarily long (finite) time

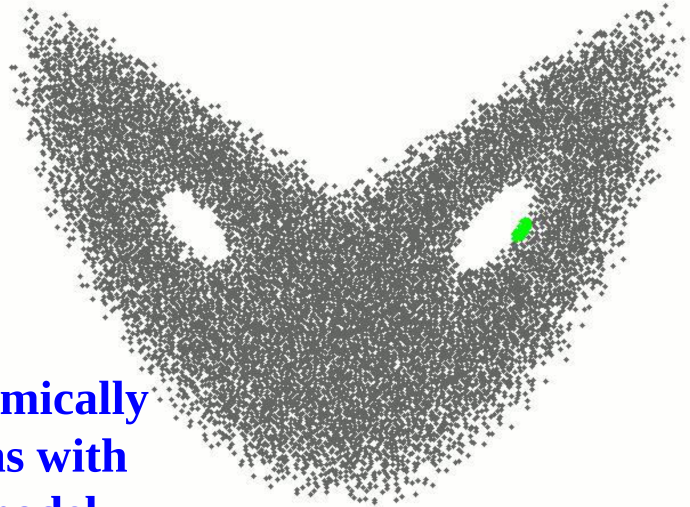


Any chance of actionable probabilities?

Predictability and Structural Model Error

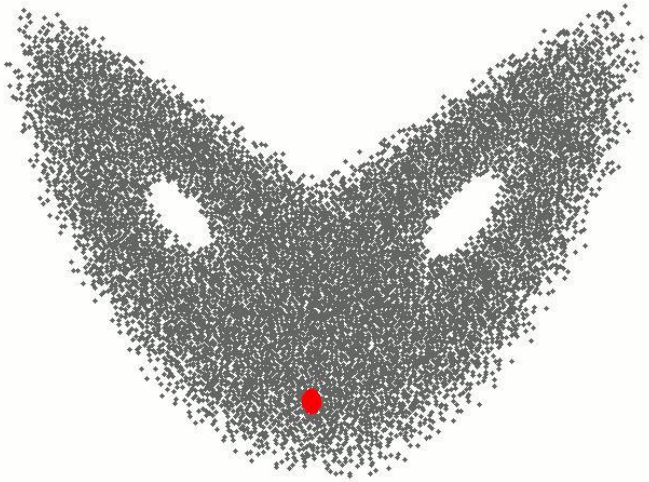
x → c sin(x/c) on RHS with c=128





An ensemble of dynamically ideal initial conditions with good but imperfect model

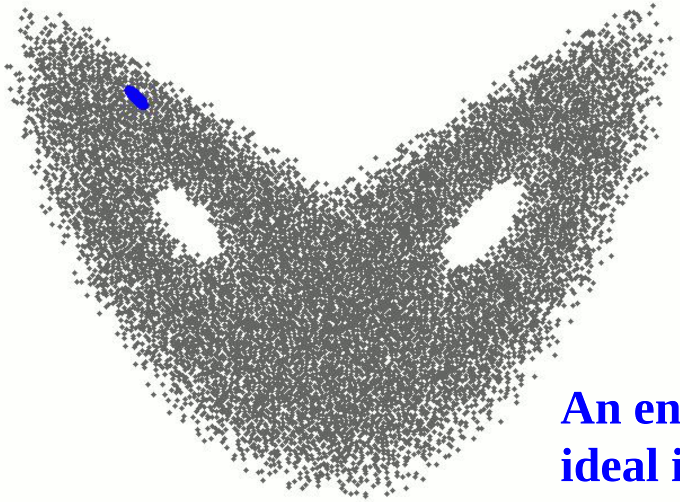
Model may shadow the system for an arbitrarily long (finite) time

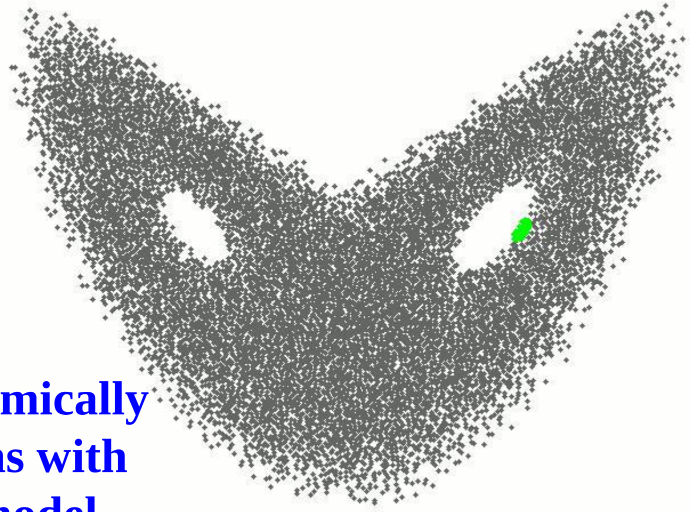


Any chance of actionable probabilities?

Predictability and Structural Model Error

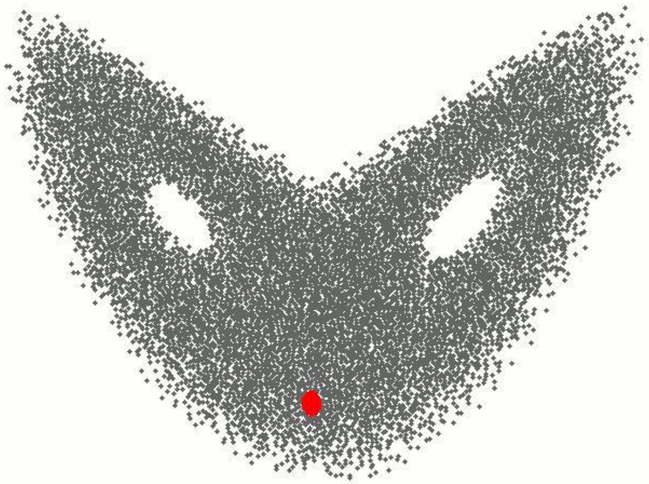
$x \rightarrow c\ \sin(x/c)$ on RHS with $c=128$





An ensemble of dynamically ideal initial conditions with good but imperfect model

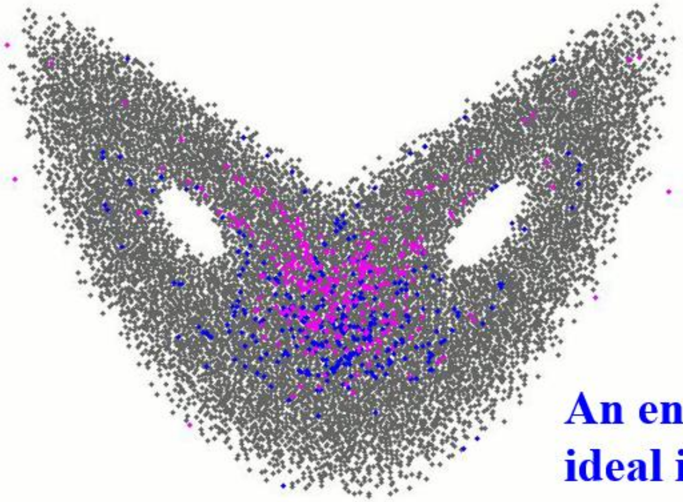
Model may shadow the system for an arbitrarily long (finite) time

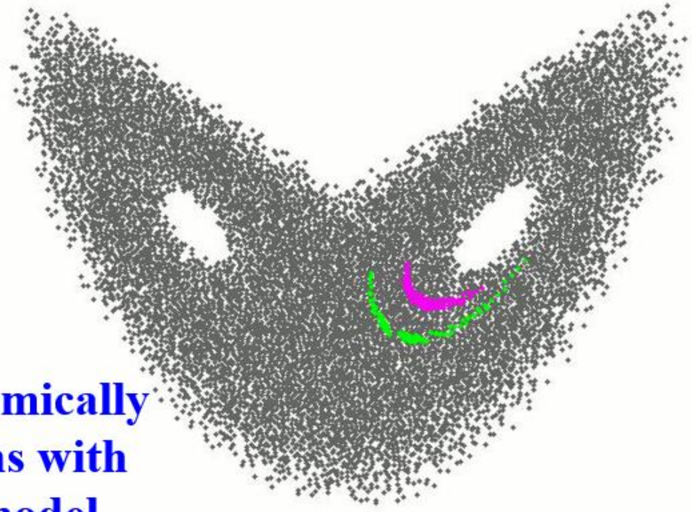


Any chance of actionable probabilities?

Predictability and Structural Model Error

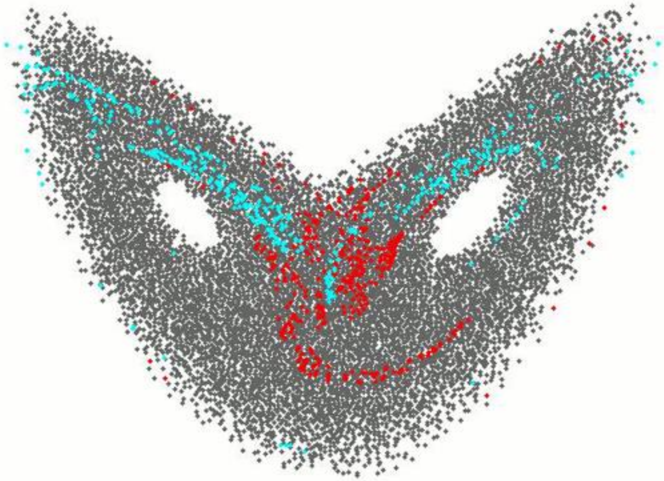
x → c sin(x/c) on RHS with c=128





An ensemble of dynamically ideal initial conditions with good but imperfect model

Model may shadow the system for an arbitrarily long (finite) time



Any chance of actionable probabilities?

The “best available” probability forecast need not be “Adequate for Purpose”

We will return to the most relevant method of measuring “skill” for a particular practitioner in a few moments.

First, note that the most skilful model to hand need not supply sufficient decisive information. Using it could in fact be disastrous.

The common Bayesian claim that one can get probabilities for everything is misguiding. Bayes can help us set up the problem correctly, it does not suggest that we can solve it.

Co-generation of tools with practitioners, may yield some that do provide enough decisive inform to aid decision making. Out of sample. This is the JEDI aim.

Forecast Direction Error
FDE for EDF

Cartoon of Problem Statement:

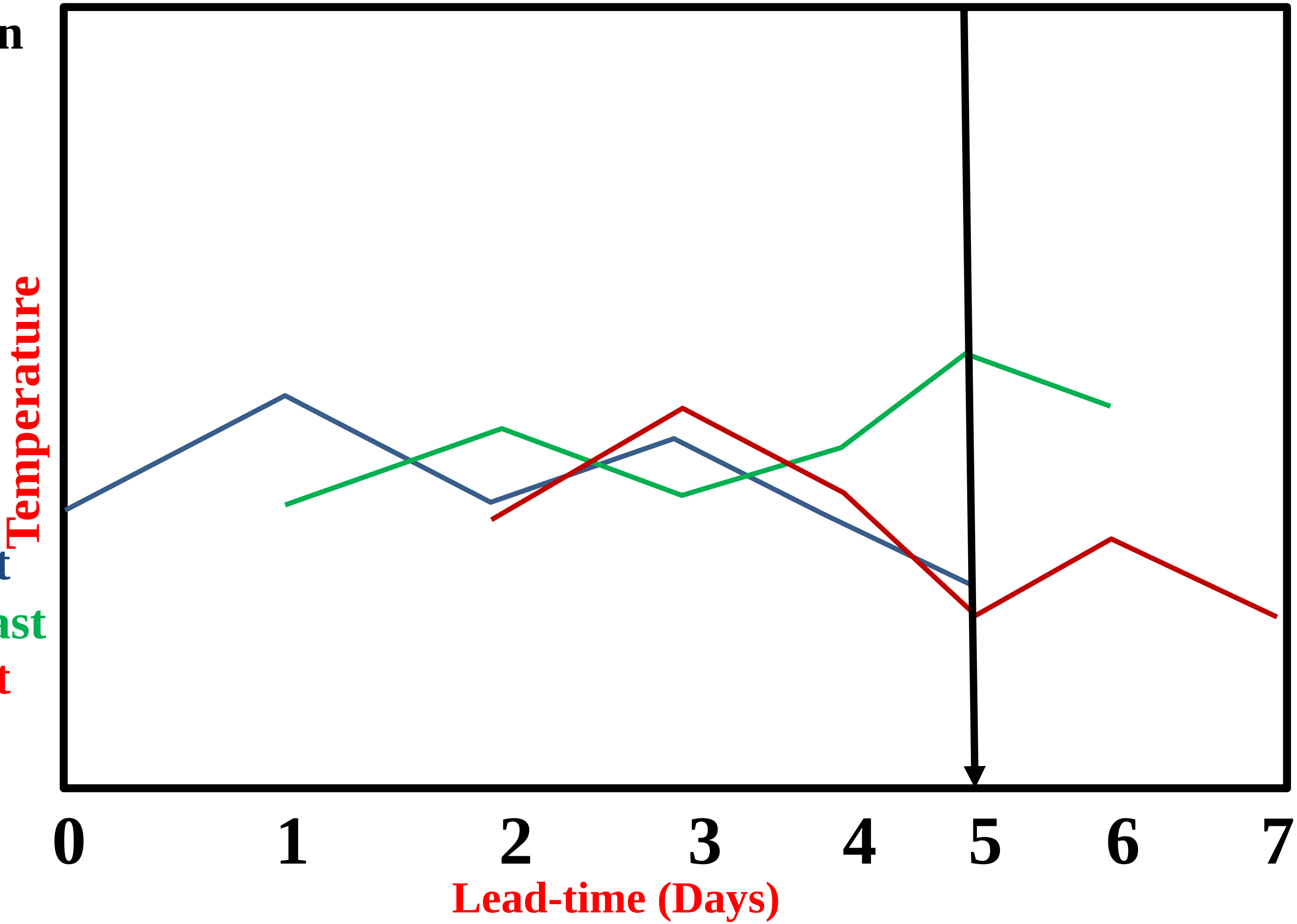
You are required by law to hold a certain amount of natural gas, the amount depends on the regulatory forecast (coloured lines). How does the forecast for Day 5 evolve?

Day 0: cold forecast

Day 1: warm forecast

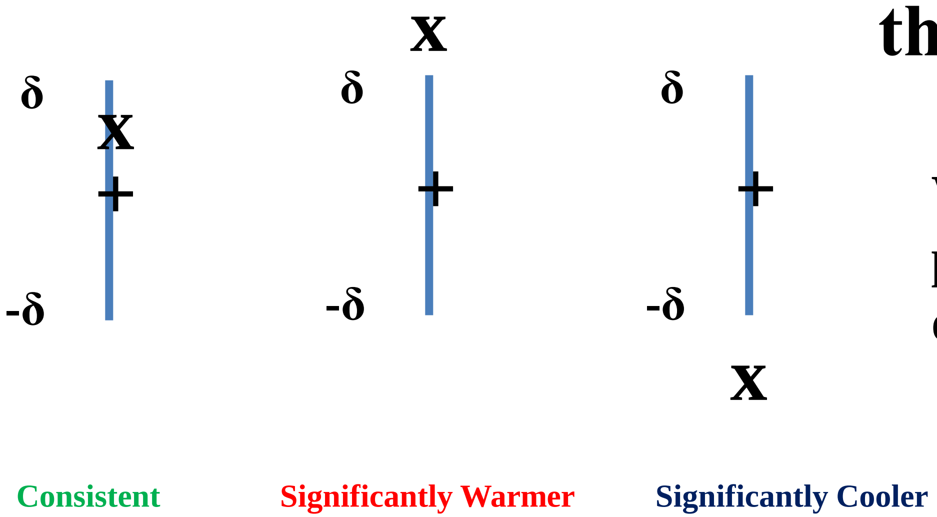
Day 2: cold forecast

Chasing the Day 5 Forecast



Forecast Direction Error
FDE for EDF: (δ, ρ)

Suppose we have the regulation model forecasts "+" the outcome is "x".



Warn the trader when the probability of exceeding a distance δ is greater than ρ.

And we could cope with small changes (< δ) in the forecast by other means.

The aim then is to spot ρ-probable forecast changes greater the δ, and ideally identify if they are positive or negative.

If we knew the true PDF of the outcome, and assumed that the regulation model was very good, this is "easy" for any δ and ρ.

Forecast Direction Error
FDE for EDF: (δ, ρ)

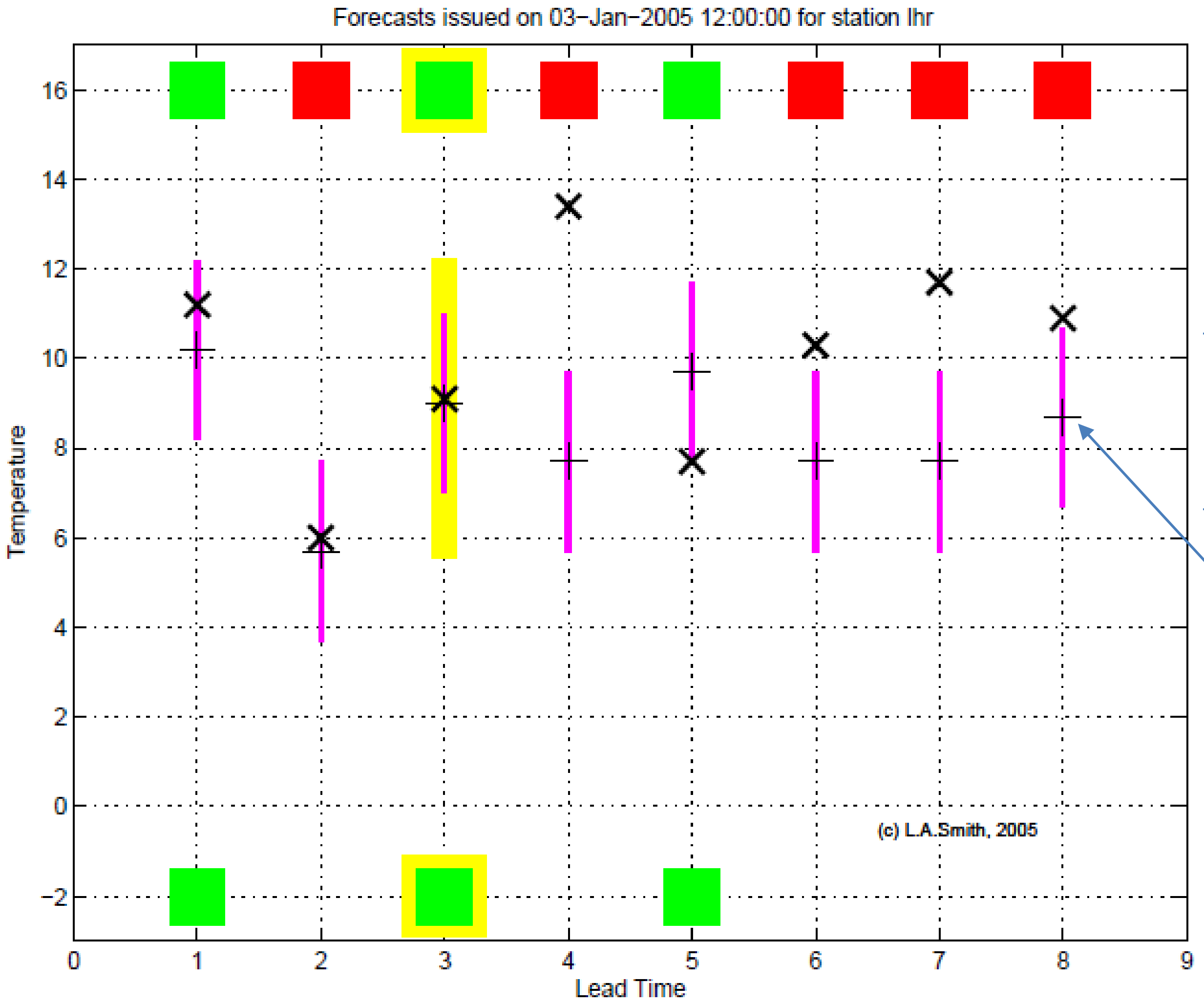
If we knew the true PDF of the outcome, and assumed that the regulation model was very good, this is "easy" for any δ and ρ.

This fails in practice!

The JEDI approach accepts this failure, and asks if there is any δ and ρ (of practical use) where the (out-of-sample) relative frequencies are consistent with a specified δ and ρ. (One must design such tests carefully.)

This worked, in real-time (truly out-of-sample) tests.

Specialised Questions
(Some answerable, some not)



Red

Green

Blue

Yellow

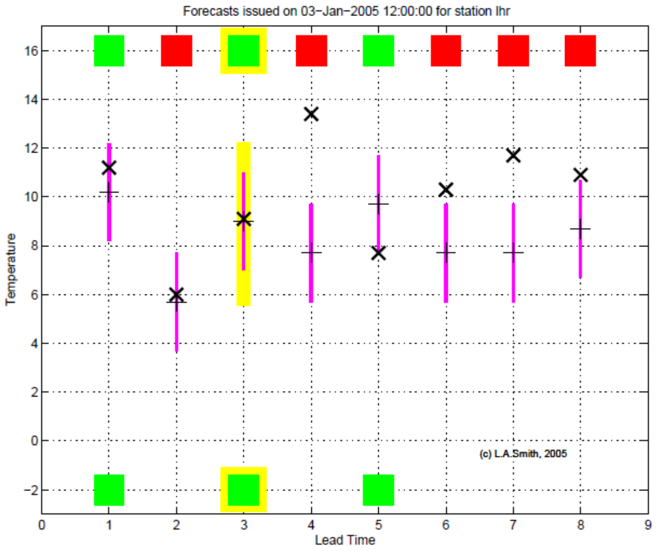
!Purple!

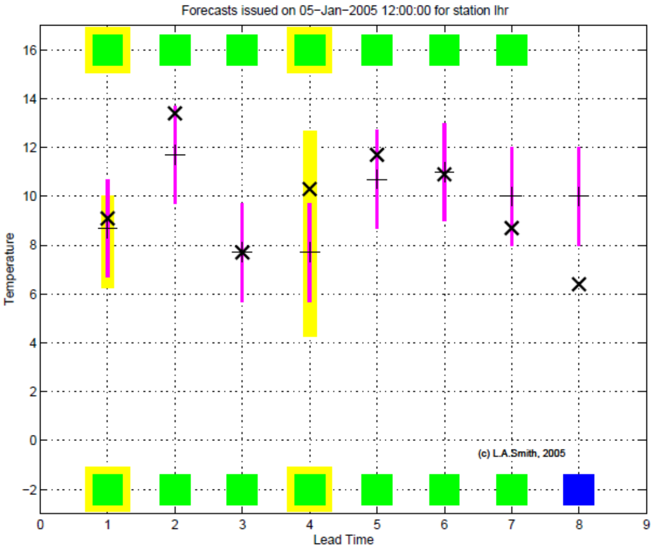
Acceptable
Range

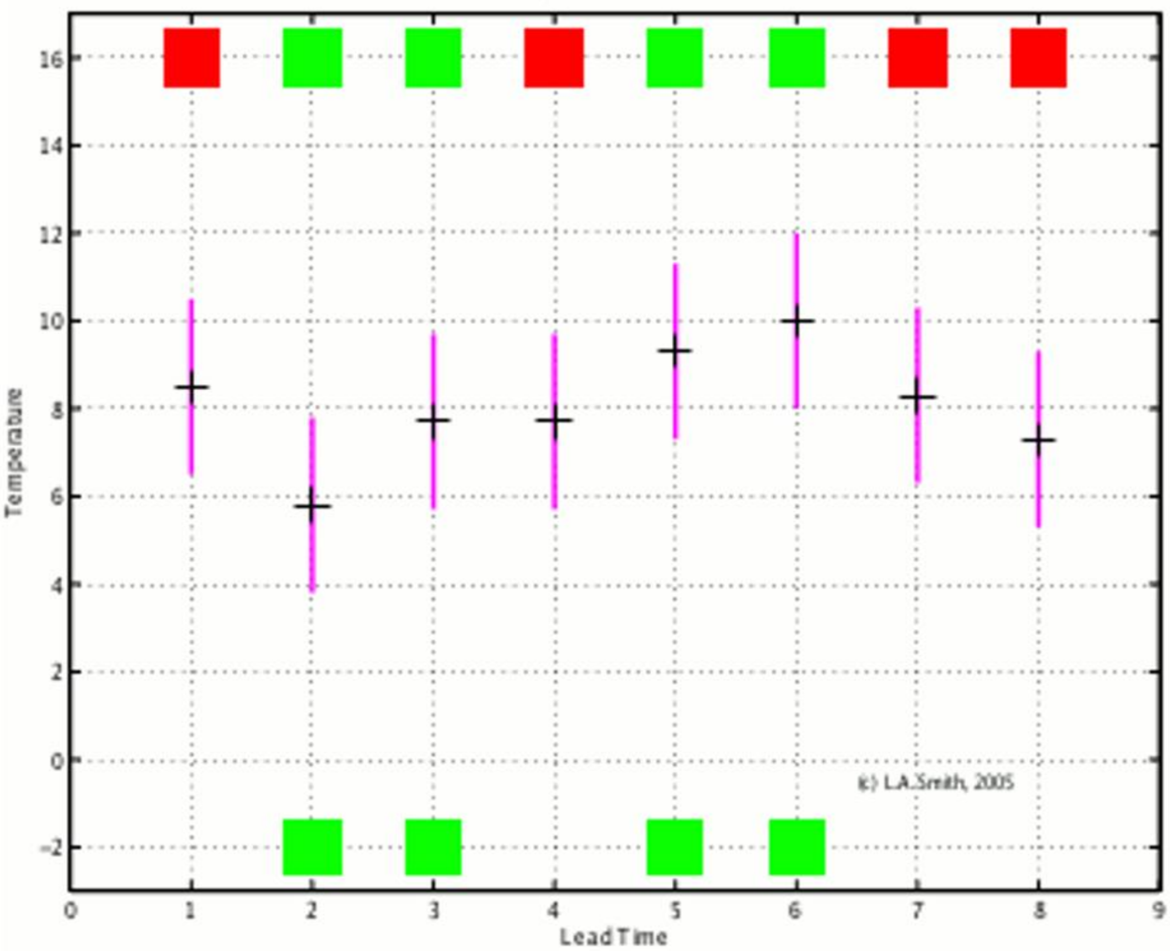
Regulatory Hi-res Forecast

Smith, L.A. (2016) 'Integrating information, misinformation and desire: improved weather-risk management for the energy sector', in Aston, P.J., Mulholland, A.J. and Tant, K.M.M. (ed.) *UK Success Stories in Industrial Mathematics*, 289-296. Springer

Some Bayesians would claim information on any threshold and tolerance could be extracted. We can not, but would welcome a year of friendly bets!



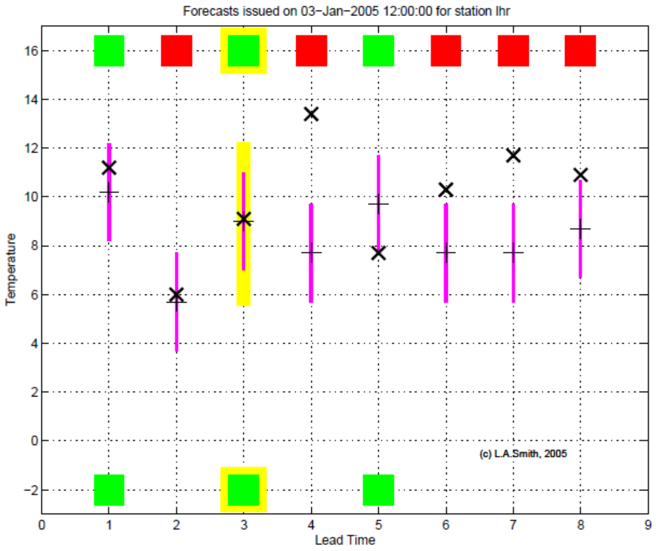


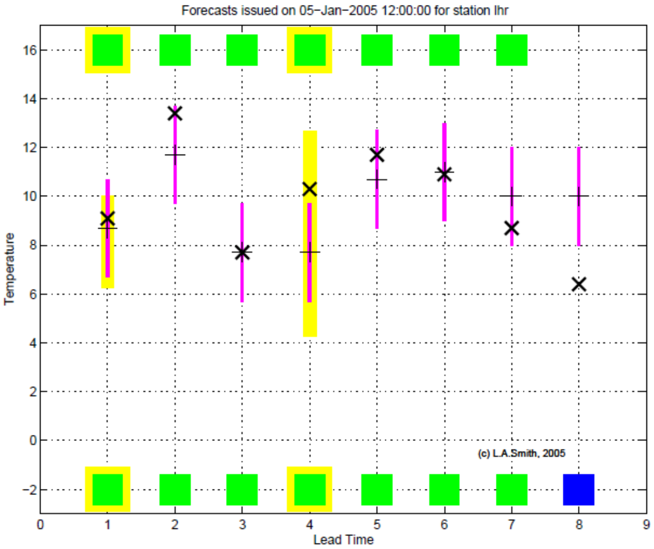


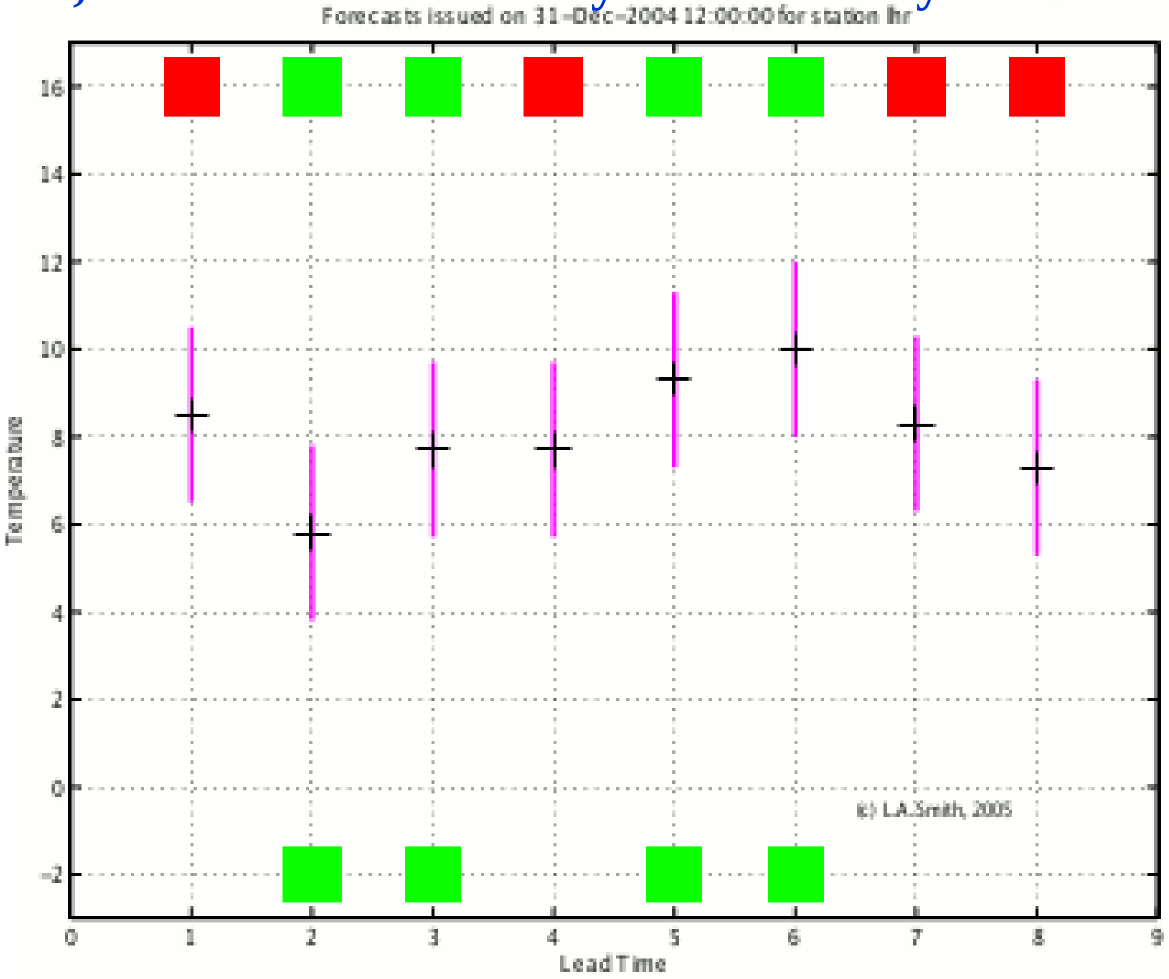
Coproduction is key!

Target needs to be doable and useful.

Some Bayesians would claim information on any threshold and tolerance could be extracted. We can not, but would welcome a year of friendly bets!

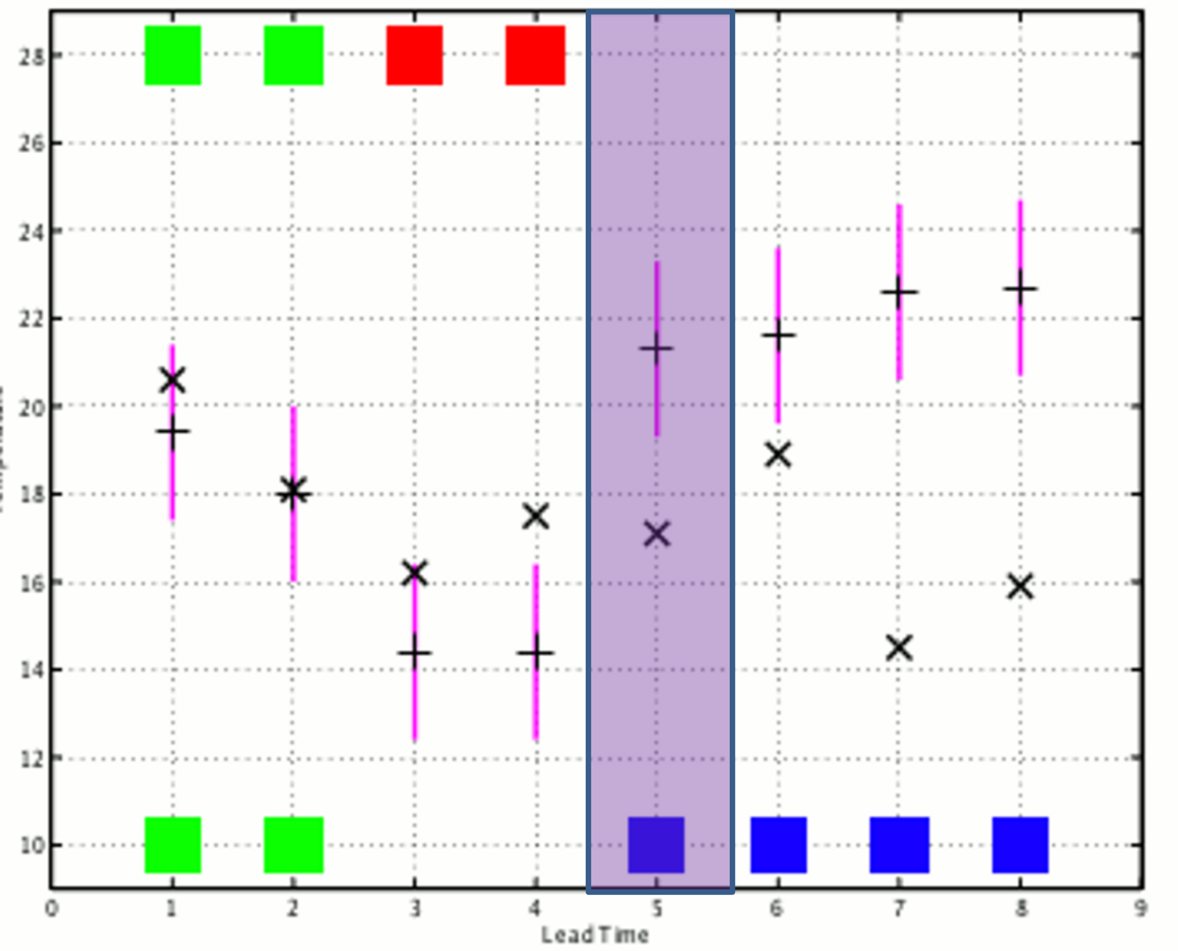






Where do the "uncertainty storms" come from?!?
They work against to aims of risk managers…
Could understanding them be of value to NWP?

Purple Light



A model which is finds itself in an unexplored (or nonsensical) region of model-state space, it issues a purple light. "Look away now."

How would an autonomous vehicle travelling at speed respond?

Forecast Direction Error
FDE for EDF

The question (always) is: Can this forecast system inform this Practitioner via this Relevant forecast about this Question?

And, of course, I treat modellers as practitioners too.
Here the question is often related to:
“How it best improve a forecast system under constraints.”

It seems silly to pretend the answer to this question is not value-laden.

Aids to Working with Practitioners Include:

Coproduction of the algorithm.

Aim for Just Enough Decisive Information (JEDI).

Adequate or Nothing (Merely Best is not sufficient)

Always include purple lights. (737)

Not Bayes Reliant, but Bayes Enabled!

Berger, J.O. and Smith, L.A. (2019) 'On the statistical formalism of uncertainty quantification' ***Annual Review of Statistics and its Application,*** **6. 3.1-3.28.**

cpt2

Different spatial models often have different levels of skill at different places. Rarely is one of them better everywhere.

This suggests assimilating the future: make pseudo-obs from each model where they are the most skilful during the forecast.

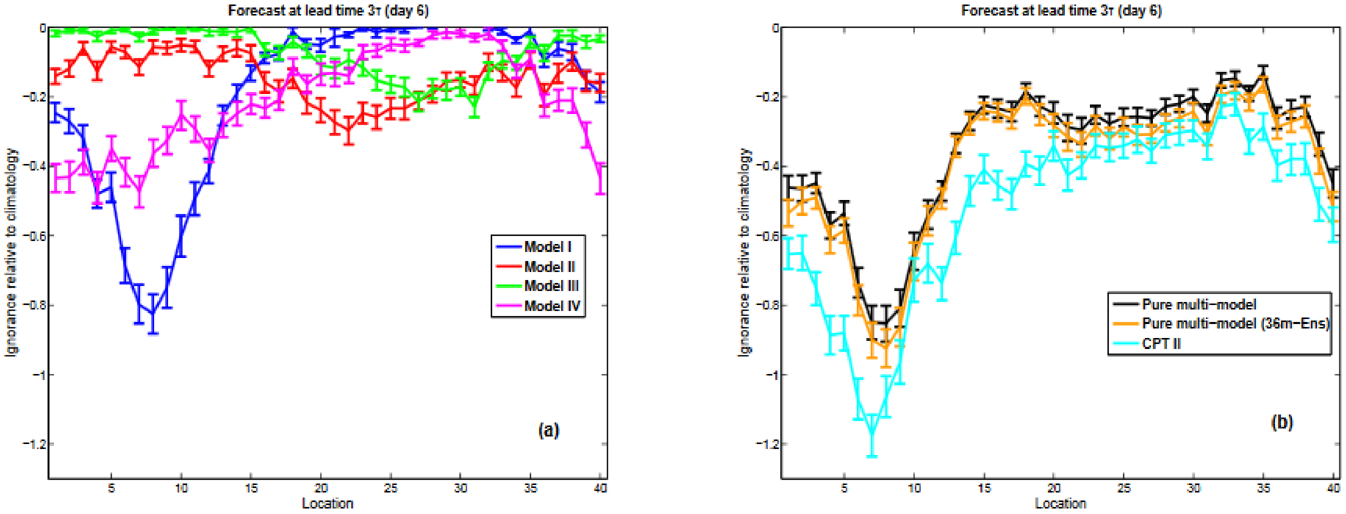
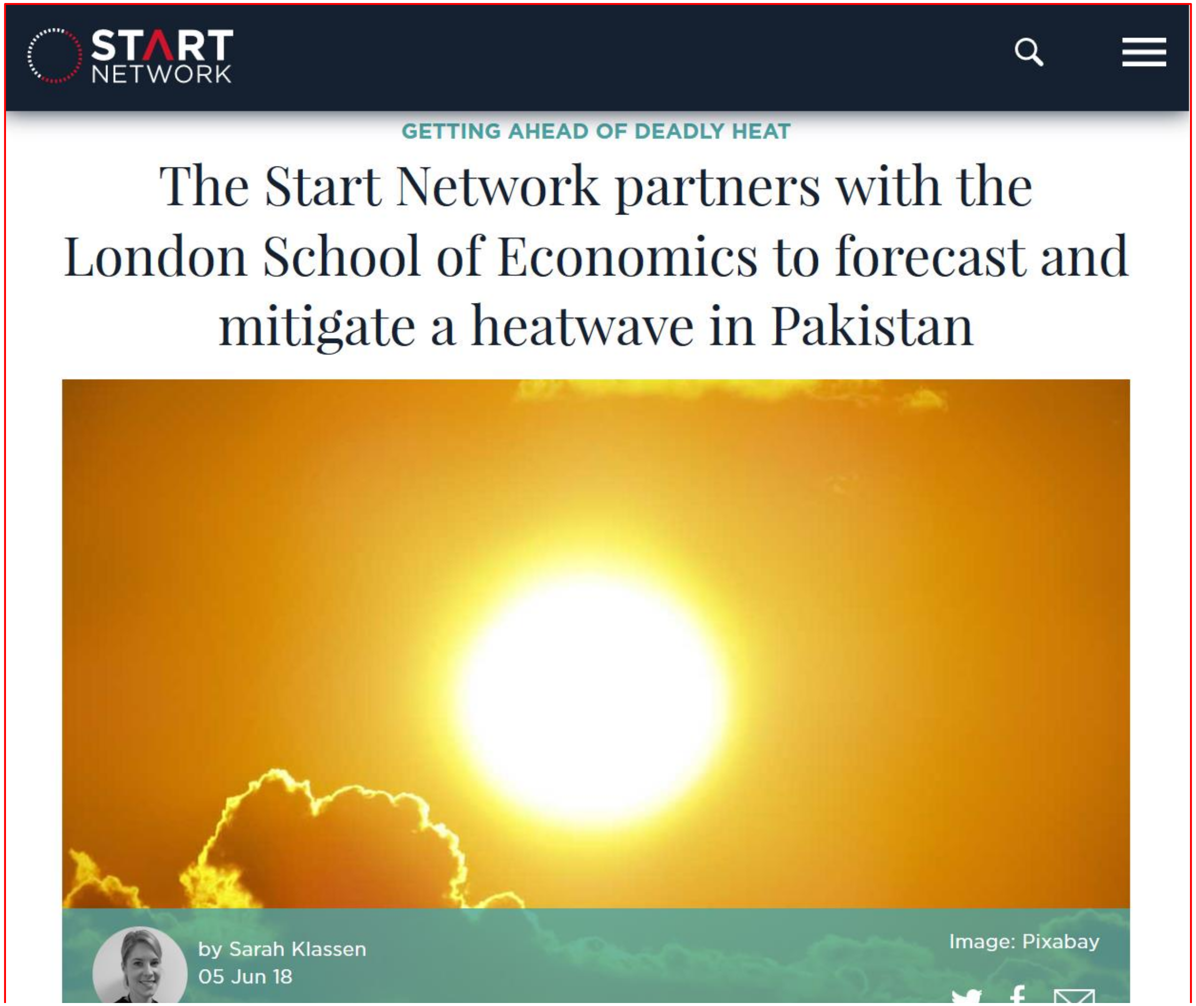


Figure 3: Ignorance score of forecasts as a function of location (model-state component) at lead-time $3\tau = 1.2$ time unit, a) forecasts from each individual model, b) pure multi-model forecast (Black), pure multi-model forecast with 36-member ensemble from each model (Brown) and CPT II forecast (Cyan).

Du, H. and Smith, L.A. (2017) 'Multimodel cross pollination in time', Physica Nonlinear Phenomena, Vol. 353-4, pp.31-38. DOI: 10.1016/j.physd.2017.06.001

Taking Forecasts off the Table (Sometimes)



GETTING AHEAD OF DEADLY HEAT

The Start Network partners with the London School of Economics to forecast and mitigate a heatwave in Pakistan

by Sarah Klassen
05 Jun 18

Image: Pixabay

Erica Thompson

In May this year, members in Pakistan raised a Start Fund alert for a heatwave, the alert was activated. Members had collectively analysed weather forecasts and had raised the alert before temperatures reached deadly levels. Start Network's Sarah Klassen discusses the challenges of forecasting heatwaves, and why a similar alert in 2017 was not activated.

https://startnetwork.org/news-and-blogs/getting-ahead-deadly-heat

Evaluating Probability Scores for the Insurance Sector
EPSIS

Sometimes a task like constructing an FDE is simply too expensive and time consuming to start off will.
In that case one would like to ask: Which Forecast System gives the best Predictive Distributions for me?

The maths I know determines how *I* want to measure skill (in my case, I J Good’s log score: IGN).

Other applied mathematicians make other choices.

But how can I learn what *you* want, without teaching you any mathematics (questionable maths a that, as all the PDFs we have to hand are imperfect!)

Evaluating Probability Scores for the Insurance Sector
EPSIS

CATS approach is to turn the question around and ask you, given two probabilistic forecasts for the same event: which one would YOU have preferred to have before the event.

We then see which (if any) of the various measures of skill reflect YOUR desires.

In the insurance sector, thus far, this inverse problem is trivial to solve: insurers tend to prefer the same distributions that Good’s Score (IGN) score as better

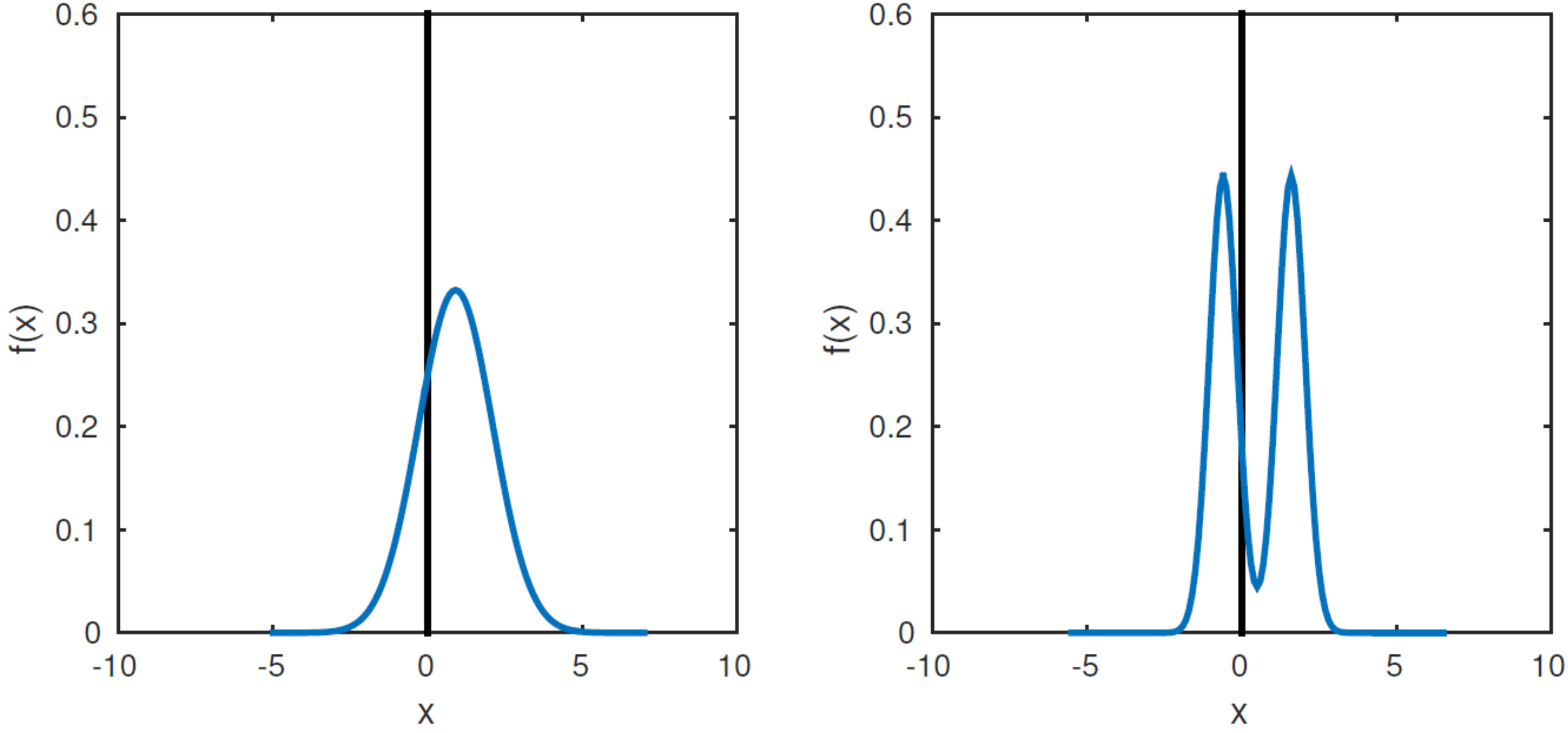
Evaluating Probability Scores for the Insurance Sector
EPSIS

If you want to help us determine what you really really want, take a look at
https://lse.eu.qualtrics.com/jfe/form/SV_bscE12V0m85bDQp

(There is a tinyurl on my last slides)

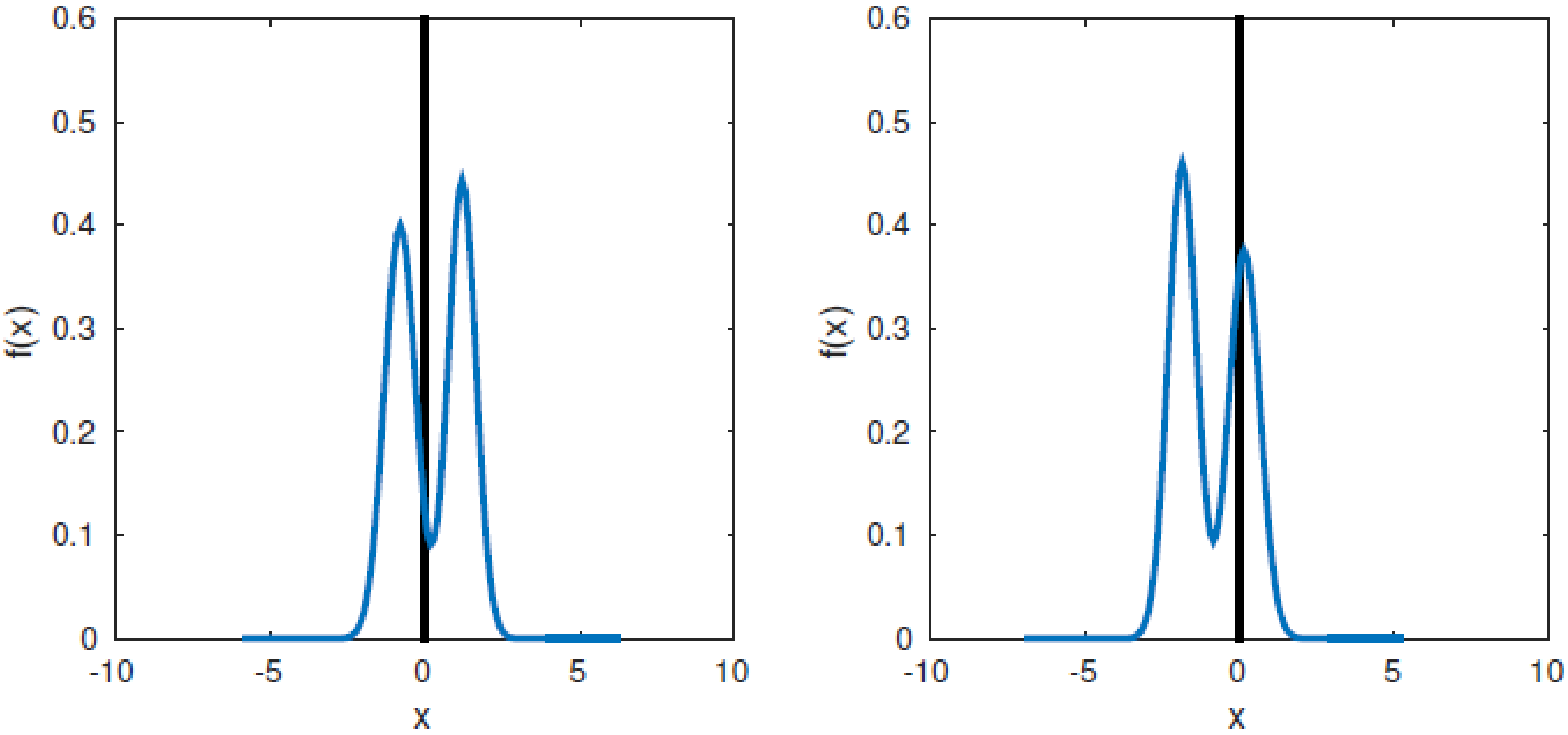
If you would like to have a copy of the EPSIS Reports at the end of the summer, please just send an email to cats@lse.ac.uk asking for one.

EPSIS

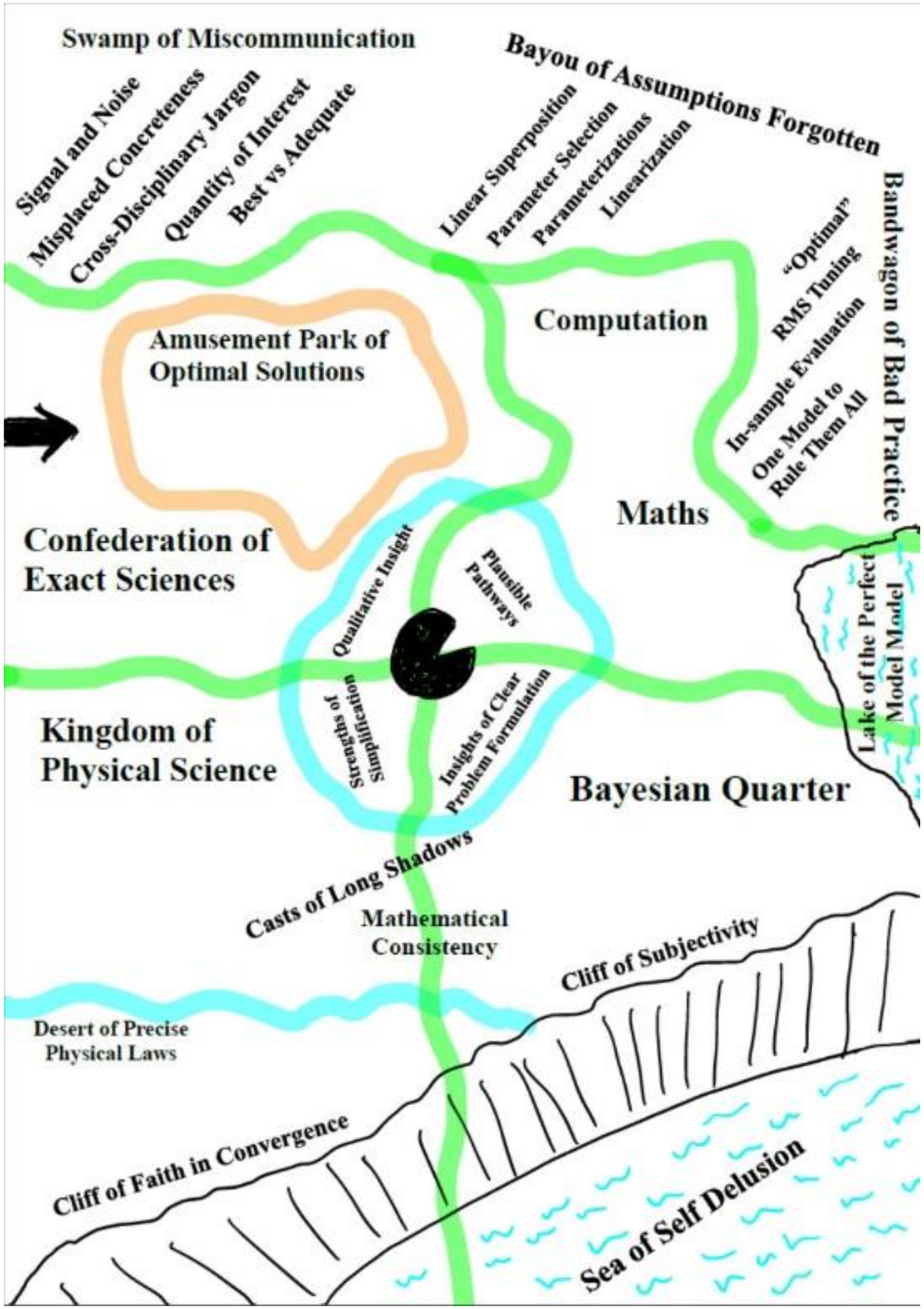


Which of these to forecast would you rather have had?

EPSIS



Which of these to forecast would you rather have had?


Swamp of Miscommunication
Signal and Noise
Misplaced Concreteness
Cross-Disciplinary Jargon
Quantity of Interest
Best vs Adequate
Bayou of Assumptions Forgotten
Linear Superposition
Parameter Selection
Parameterizations
Linearization
Bandwagon of Bad Practice
"Optimal"
RMS Tuning
In-sample Evaluation
One Model to Rule Them All
Computation
Amusement Park of Optimal Solutions
Maths
Confederation of Exact Sciences
Qualitative Insight
Plausible Pathways
Lake of the Perfect Model Model
Kingdom of Physical Science
Strengths of Simplification
Insights of Clear Problem Formulation
Bayesian Quarter
Casts of Long Shadows
Mathematical Consistency
Cliff of Subjectivity
Desert of Precise Physical Laws
Cliff of Faith in Convergence
Sea of Self Delusion

QUESTIONS?

?ANSWERS?

Tinyurl.com/y67dm9oo To select your PDFs.

Slido.com #D571 To ask questions

@lynyrdsmyth @H4wkm0th To follow the conversation

References

CATS@lse.ac.uk

J Berger and LA Smith (2018) Uncertainty Quantification, Annual Reviews of Statistics (to appear). **Annual Review of Statistics and Its Application**
Vol. 6:433-460 (Volume publication date March 2019)

Smith, L.A. (1995) 'Accountability and error in ensemble forecasting', In 1995 ECMWF Seminar on Predictability. Vol. 1, 351-368. ECMWF, Reading.

Smith, L.A. (2016) 'Integrating information, misinformation and desire: improved weather-risk management for the energy sector', in Aston, P et al. (ed.) *UK Success Stories in Industrial Mathematics*, 289-296. Springer

Du, H. & Smith, L.A. (2017) Multi-model cross-pollination in time Physica D 353, p. 31-38.

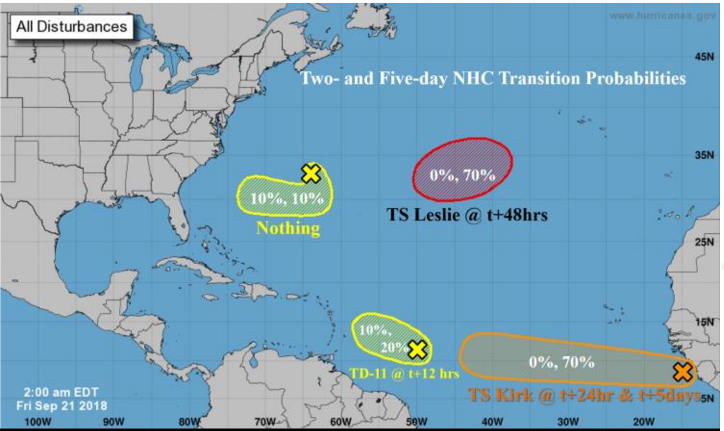
K Judd, CA Reynolds, LA Smith & TE Rosmond (2008) The Geometry of Model Error. JAS 65 (6), 1749-1772.

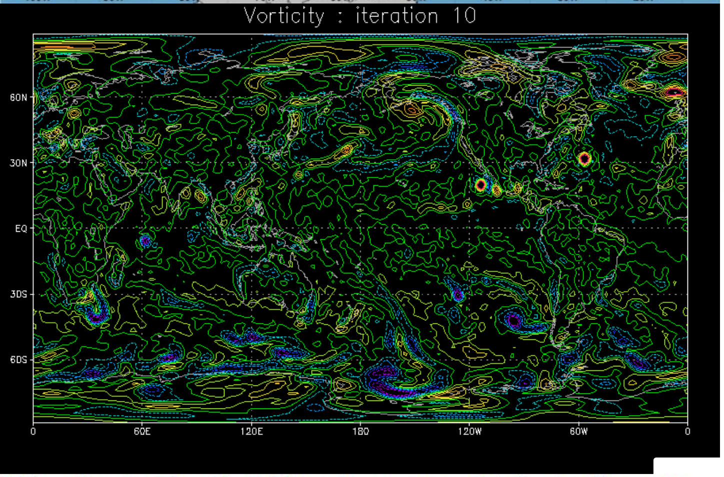
Tinyurl.com/y67dm9oo To select your PDFs.

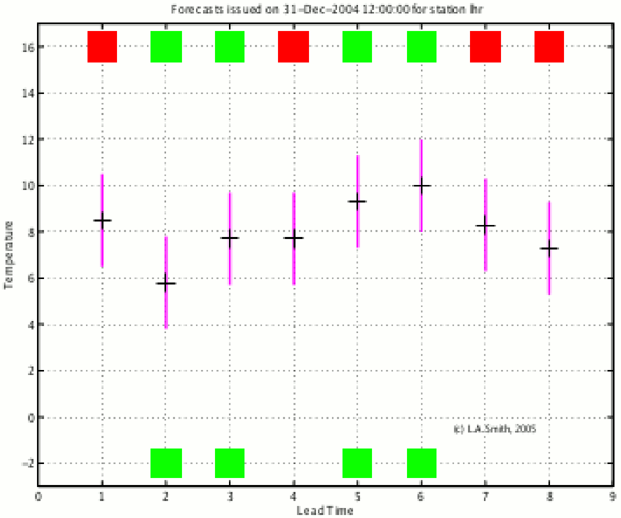
Slido.com #D571 To ask questions

@lynyrdsmyth @H4wkm0th To follow the conversation

END







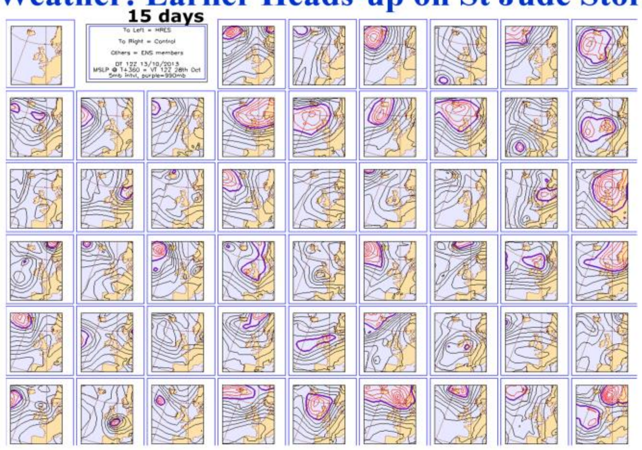
Applications of our approach are widespread

FDE Electricity Demand for EDF

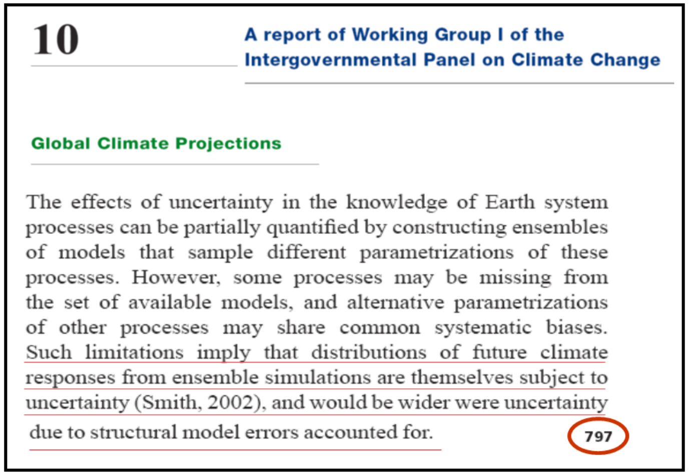
Hurricane Guidance	Nuclear Power
Data Assimilation	Hunting Licences
Ensemble Weather	RNLI Guidance
	Nuclear Stewardship

The IPCC acknowledges implications of working in model land explicitly.

Weather: Earlier Heads-up on St Jude Storm



Thanks to Tim Hewson and ECMWF



10

A report of Working Group I of the Intergovernmental Panel on Climate Change

Global Climate Projections

The effects of uncertainty in the knowledge of Earth system processes can be partially quantified by constructing ensembles of models that sample different parametrizations of these processes. However, some processes may be missing from the set of available models, and alternative parametrizations of other processes may share common systematic biases. Such limitations imply that distributions of future climate responses from ensemble simulations are themselves subject to uncertainty (Smith, 2002), and would be wider were uncertainty due to structural model errors accounted for.

797

Nuclear Stewartship

(by dropping balls off towers)

Ball(s)

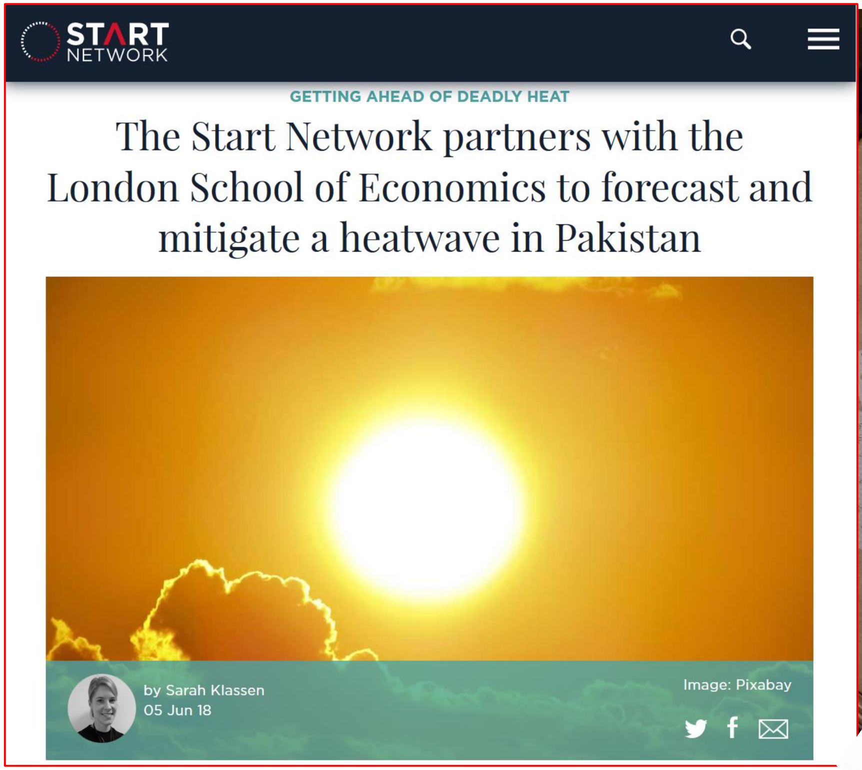
1000 ft Ua1 shaft

2 bowling balls
3 Basketballs
2 golf balls
3 Wiffle balls
… (no rubber duck)

ors

http://www2.nstec.com/Documents/Fact%20Sheets/U1a%20Facility.pdf

Real Forecasts are focused on a Question



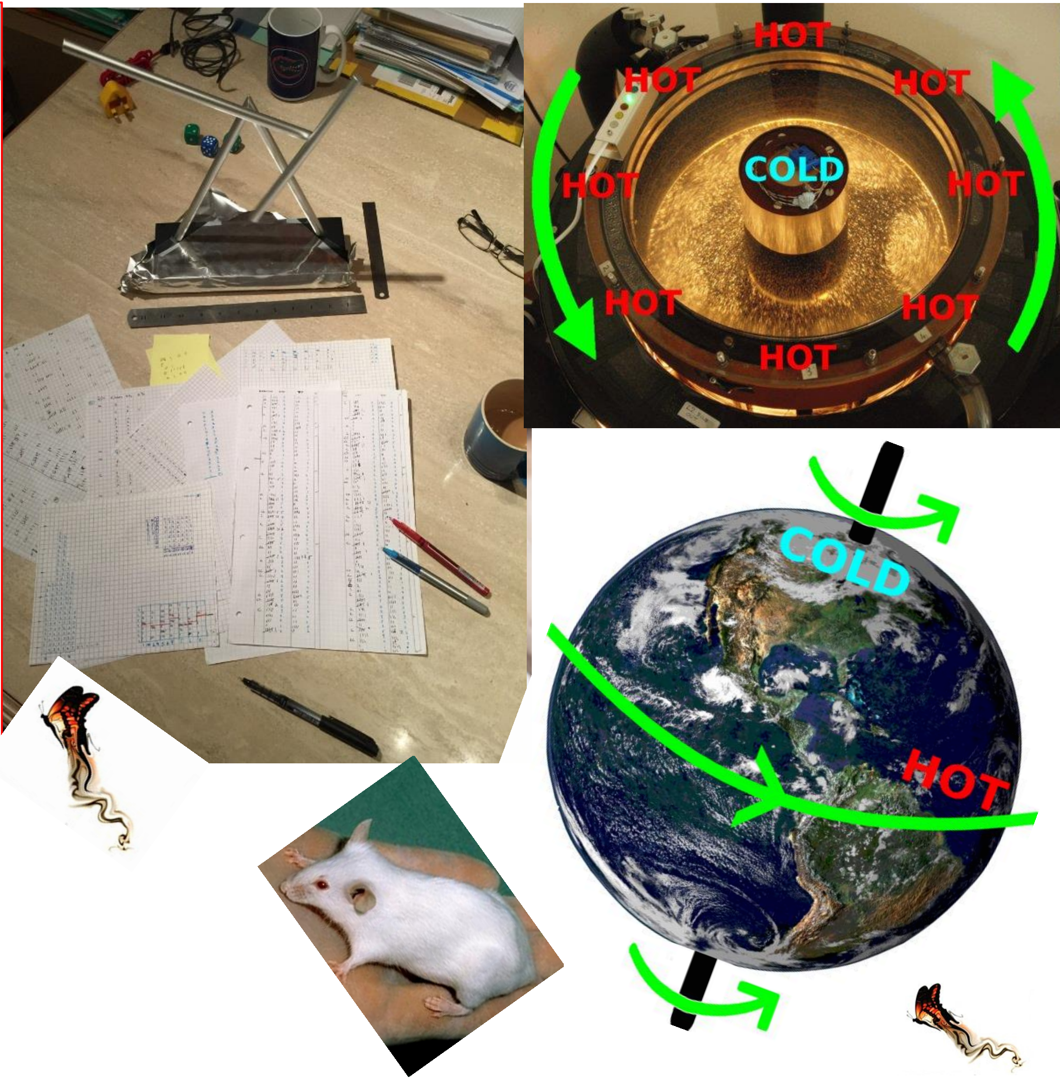
START NETWORK

GETTING AHEAD OF DEADLY HEAT

The Start Network partners with the London School of Economics to forecast and mitigate a heatwave in Pakistan

by Sarah Klassen
05 Jun 18

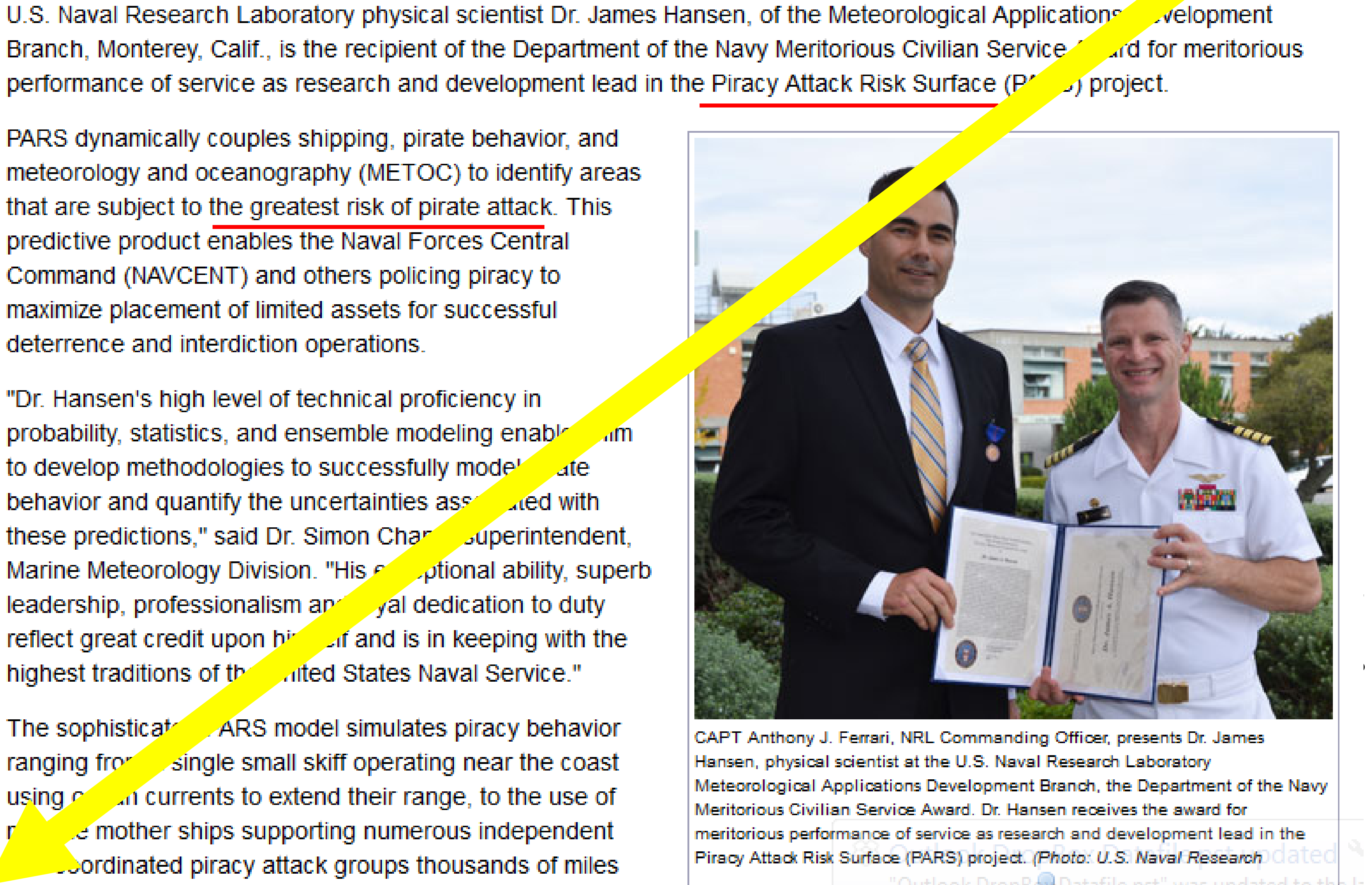
Image: Pixabay



What is Model-land?

Note in passing that not all models are mathematical. ?Analog UQ?

Things are NOT HOPELESS (Useless)!
A Weather-like task: Predicting Pirates



U.S. Naval Research Laboratory physical scientist Dr. James Hansen, of the Meteorological Applications Development Branch, Monterey, Calif., is the recipient of the Department of the Navy Meritorious Civilian Service Award for meritorious performance of service as research and development lead in the Piracy Attack Risk Surface (PARS) project.

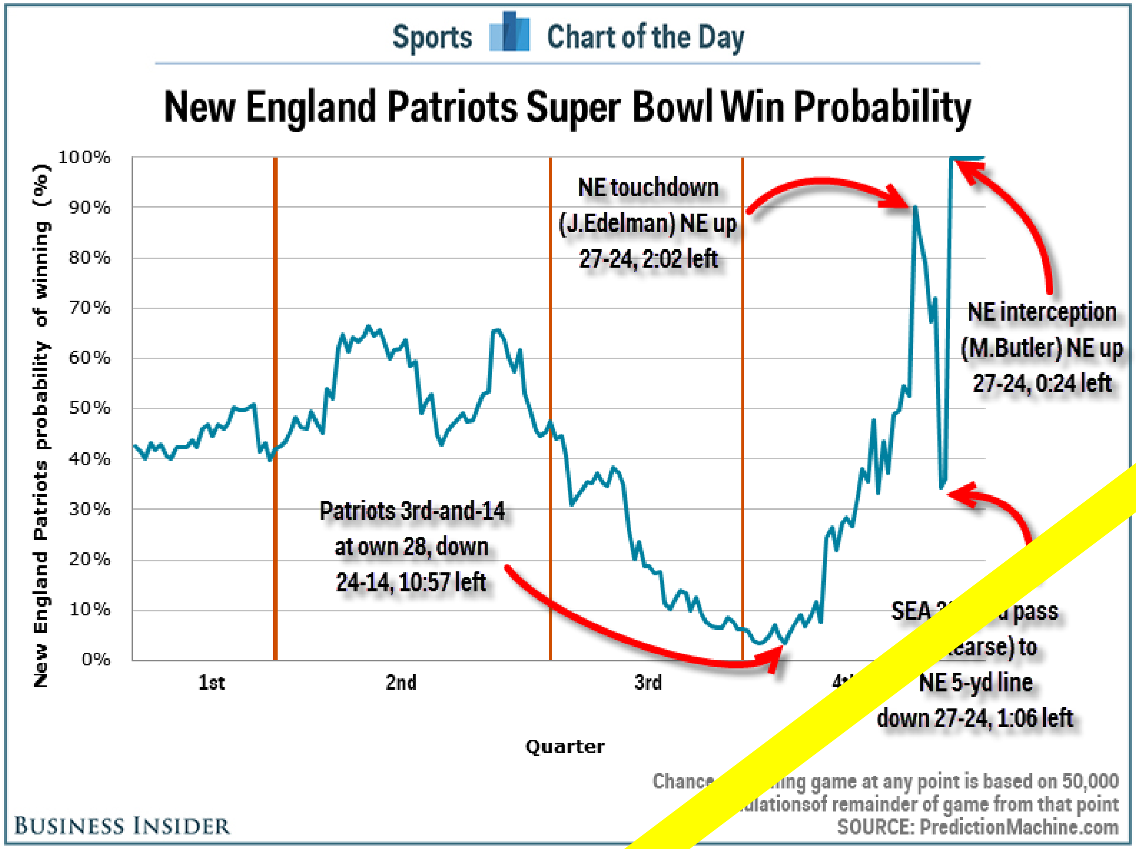
PARS dynamically couples shipping, pirate behavior, and meteorology and oceanography (METOC) to identify areas that are subject to the greatest risk of pirate attack. This predictive product enables the Naval Forces Central Command (NAVCENT) and others policing piracy to maximize placement of limited assets for successful deterrence and interdiction operations.

"Dr. Hansen's high level of technical proficiency in probability, statistics, and ensemble modeling enabled him to develop methodologies to successfully model pirate behavior and quantify the uncertainties associated with these predictions," said Dr. Simon Chang, superintendent, Marine Meteorology Division. "His exceptional ability, superb leadership, professionalism and loyal dedication to duty reflect great credit upon himself and is in keeping with the highest traditions of the United States Naval Service."

The sophisticated PARS model simulates piracy behavior ranging from a single small skiff operating near the coast using ocean currents to extend their range, to the use of multiple mother ships supporting numerous independent or coordinated piracy attack groups thousands of miles

CAPT Anthony J. Ferrari, NRL Commanding Officer, presents Dr. James Hansen, physical scientist at the U.S. Naval Research Laboratory Meteorological Applications Development Branch, the Department of the Navy Meritorious Civilian Service Award. Dr. Hansen receives the award for meritorious performance of service as research and development lead in the Piracy Attack Risk Surface (PARS) project. *(Photo: U.S. Naval Research*

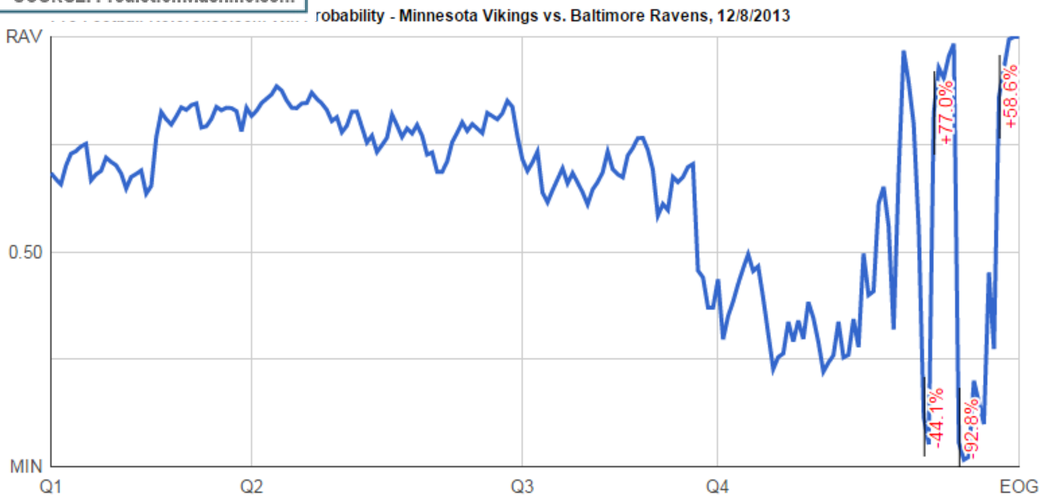
Probability of Success after start of a Mission



What is the orrect way to ake evolving prob bilities?

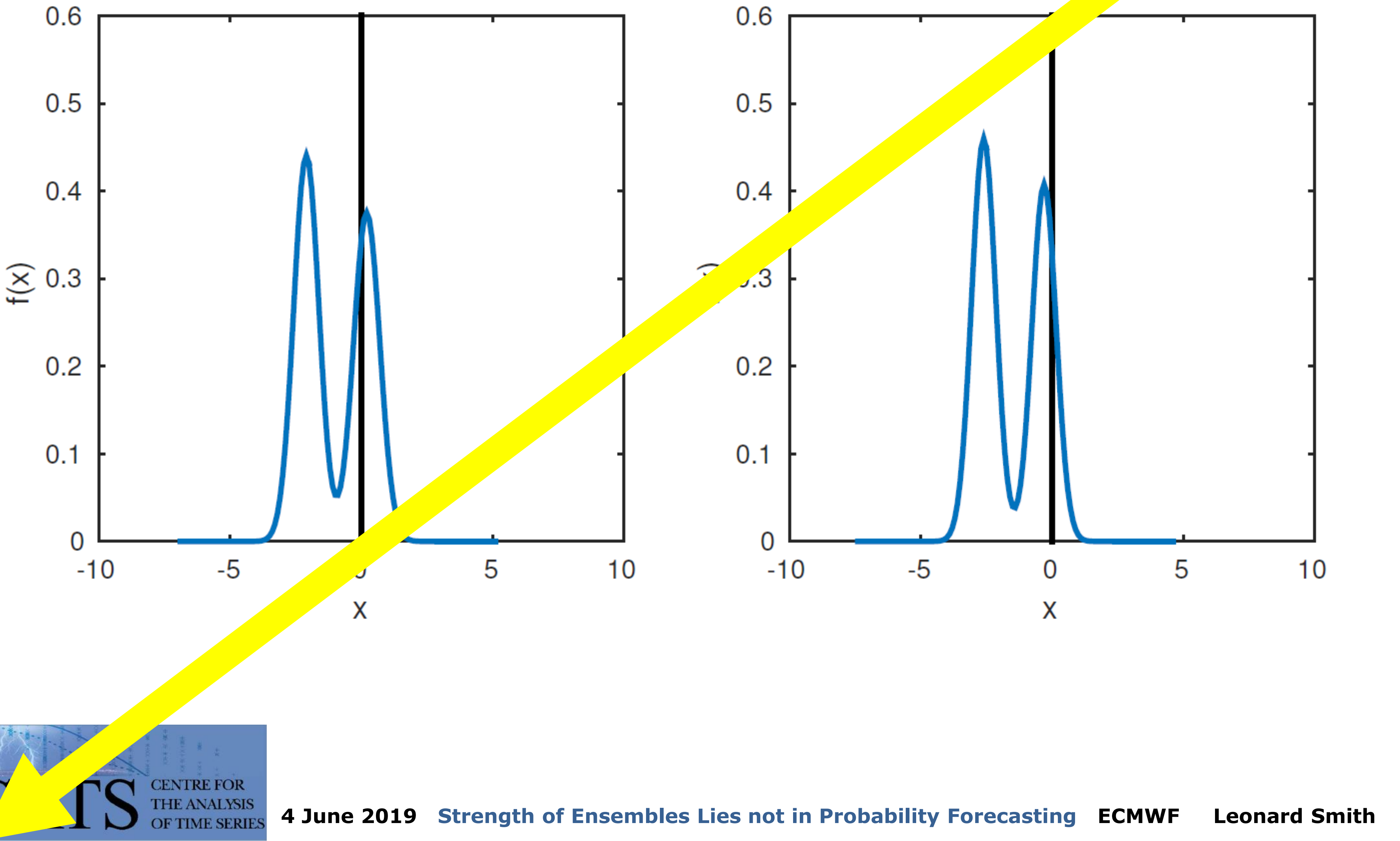
ow can we evaluate this kind of forecast system?

I do not know how to this correctly. Taking the “best” each point in time is not enough

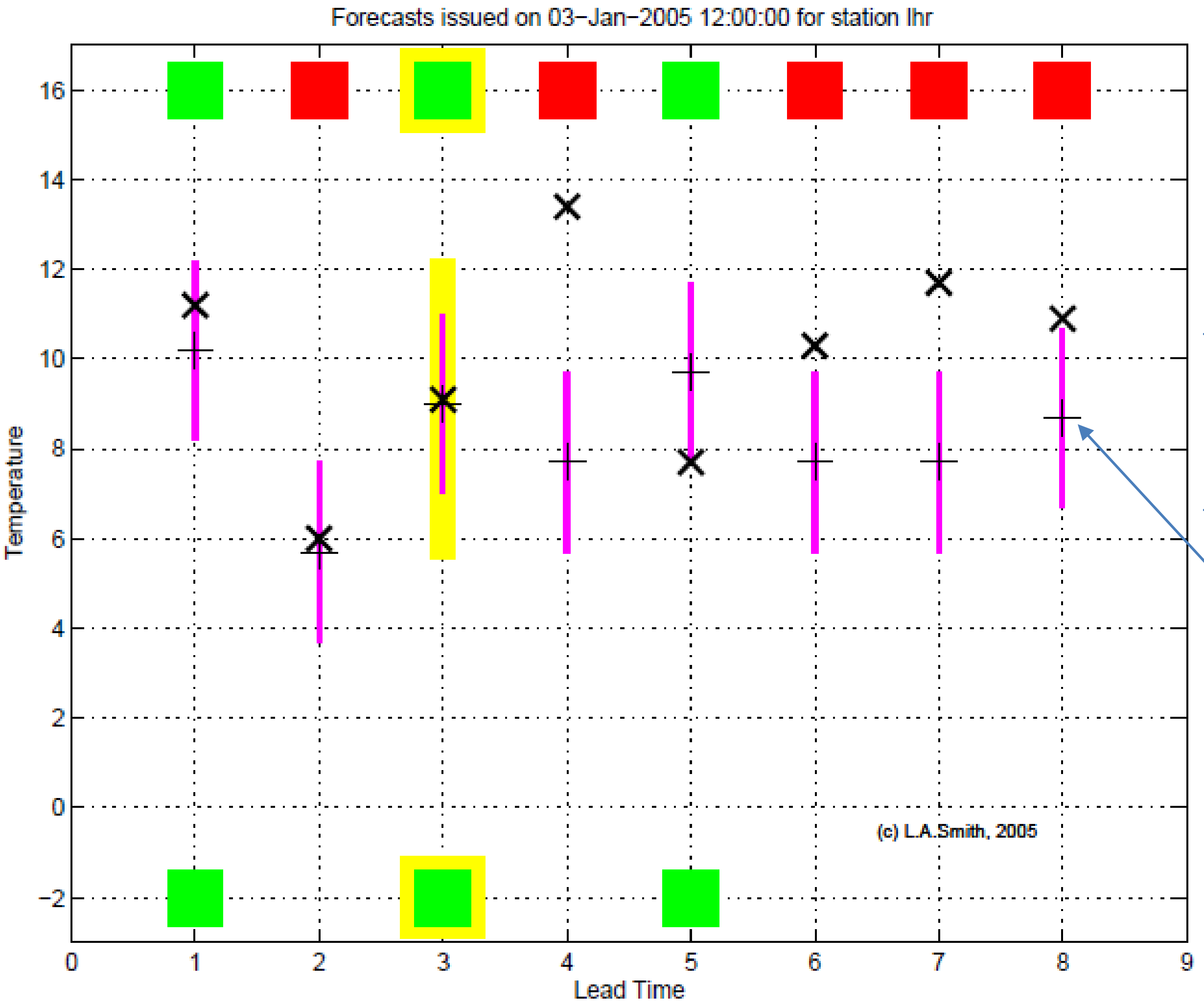


CENTRE FOR THE ANALYSIS OF TIME SERIES

EPSIS



Specialised Questions
(Some answerable, some not)



Red
Green
Blue

Yellow

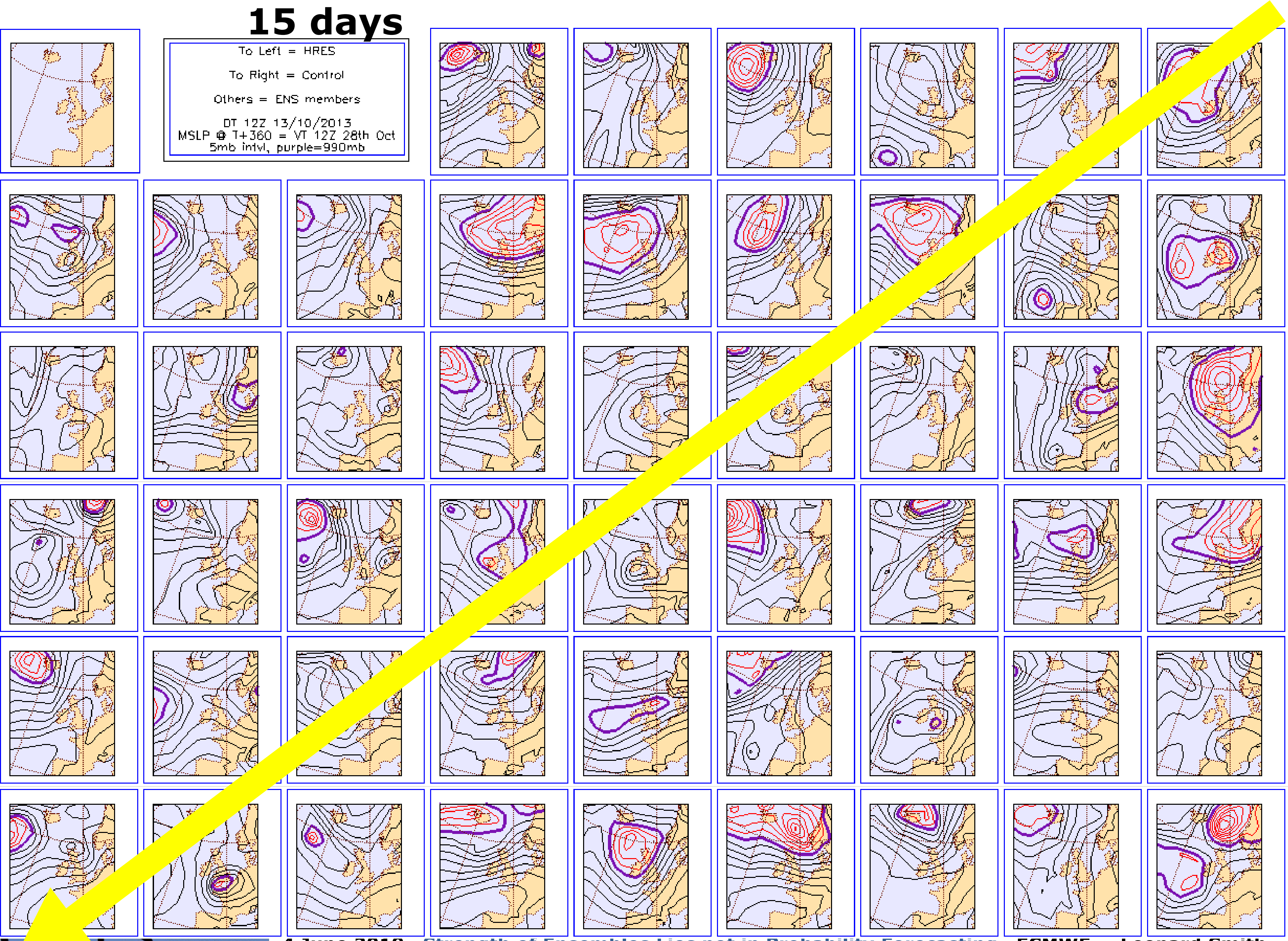
!Purple!

Acceptable Range

Regulatory Hi-res Forecast

Smith, L.A. (2016) '[Integrating information, misinformation and desire: improved weather-risk management for the energy sector](#)', in Aston, P.J., Mulholland, A.J. and Tant, K.M.M. (ed.) *UK Success Stories in Industrial Mathematics*, 289-296. Springer

15 days



To Left = HRES

To Right = Control

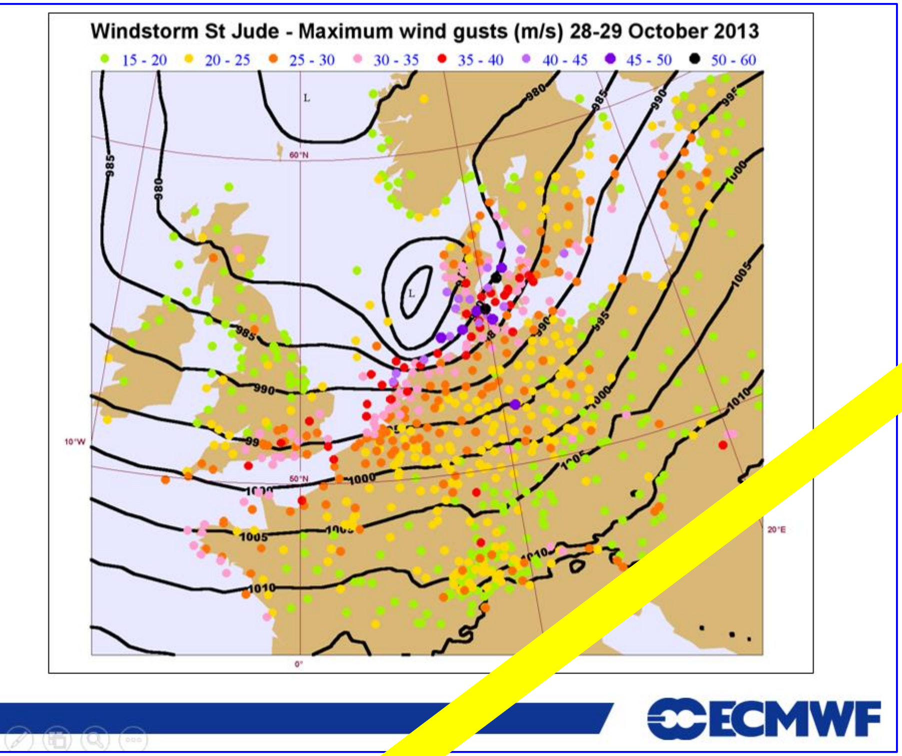
Others = ENS members

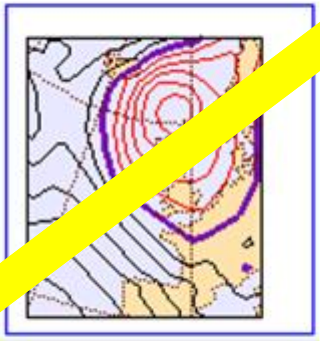
DT 12Z 13/10/2013
MSLP @ T+360 = VT 12Z 28th Oct
5mb intvl, purple=990mb

Real-world Targets: Getting out of Model-land

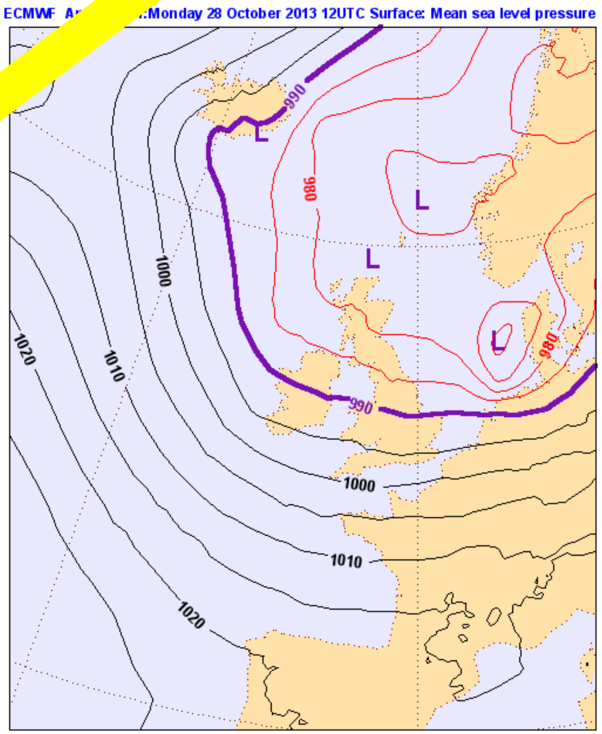
Observations of the storm

and the ECMWF analysis





T-15 days


ECMWF An :Monday 28 October 2013 12UTC Surface: Mean sea level pressure
990
980
1000
1010
1020
L

Today’s models provide sufficiently good forward simulation that neither chaos nor model error make the ensemble useless even in week two!

Thanks to ECMWF &Tim Hewson

That does not, of course, imply we can extract useful probabilities.

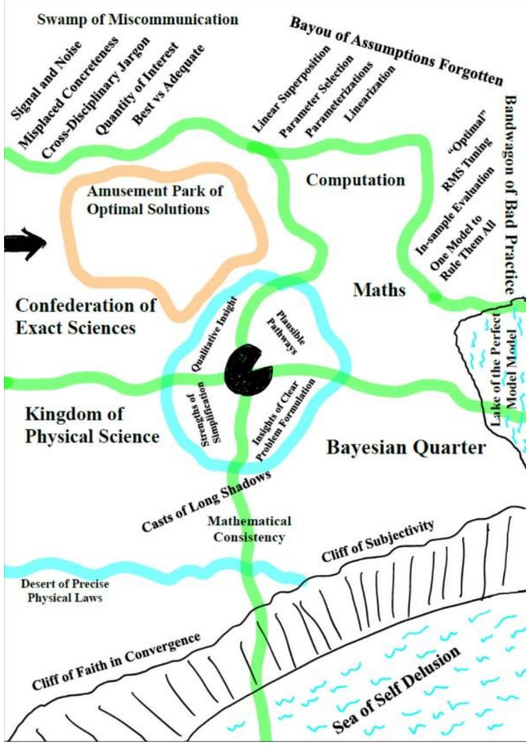
CATS CENTRE FOR THE ANALYSIS OF TIME SERIES

Purple Lights and Probabilities

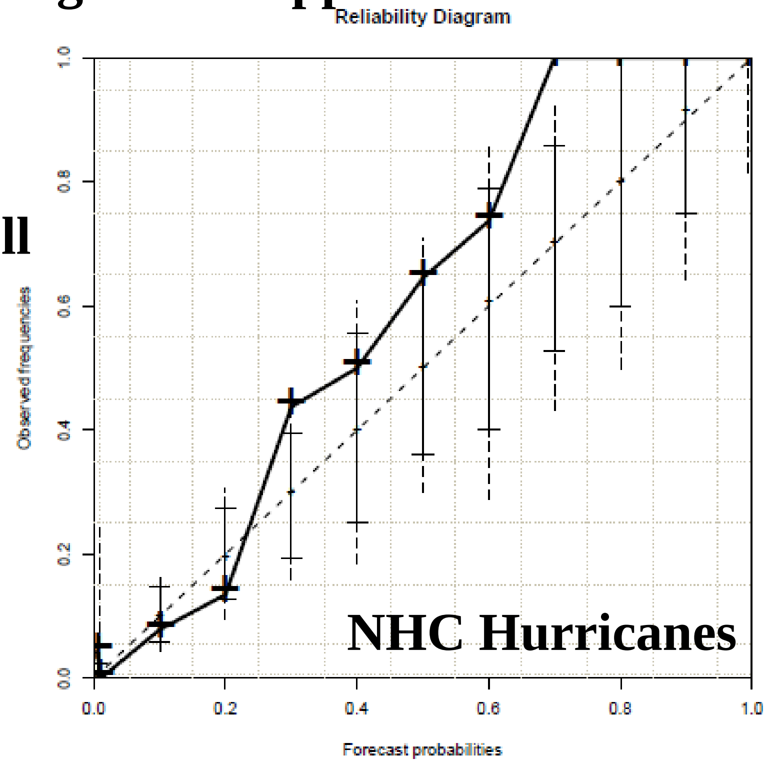
What "probability" should you offer given a purple light?

What probability should you offer if your predicted probabilities are inconsistent with the observed relative frequencies?

What probability should you offer when something (previously) unimaginable happens?


Swamp of Miscommunication
Signal and Noise
Misplaced Concreteness
Cross-Disciplinary Jargon
Quantity of Interest
Best vs Adequate
Bayou of Assumptions Forgotten
Linear Superposition
Parameter Selection
Parameterizations
Linearization
Bandwagon of Bad Practice
"Optimal"
RMS Tuning
In-sample Evaluation
One Model to Rule Them All
Computation
Amusement Park of Optimal Solutions
Maths
Confederation of Exact Sciences
Qualitative Insight
Plausible Pathways
Lake of the Perfect Model Model
Kingdom of Physical Science
Strengths of Simplification
Insights of Clear Problem Formulation
Bayesian Quarter
Casts of Long Shadows
Mathematical Consistency
Cliff of Subjectivity
Desert of Precise Physical Laws
Cliff of Faith in Convergence
Sea of Self Delusion

What will you tell an autonomous vehicle to do?



NHC Hurricanes

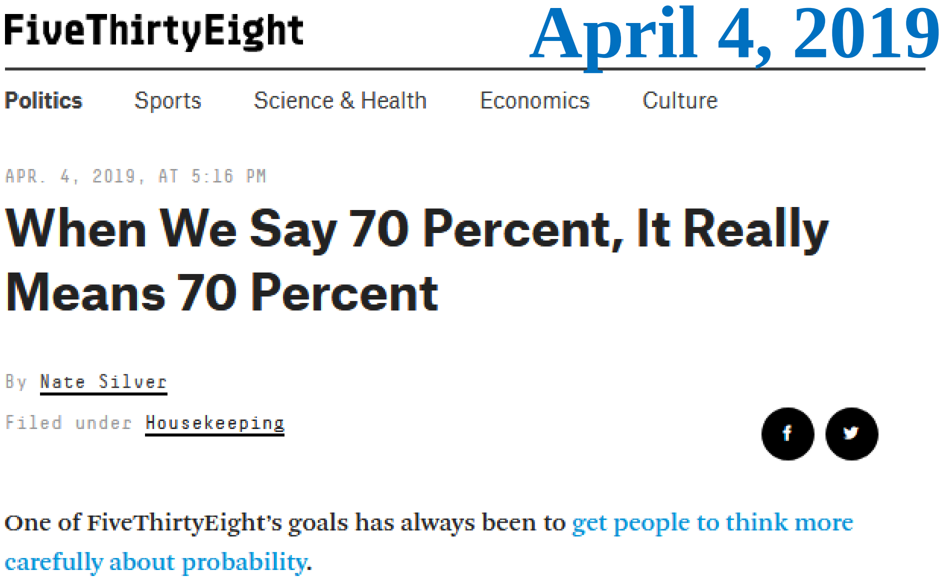
Figure 6.2: NHC 2012 TC forecast reliability: reliability diagram for the NHC's 2012 48-hr TC forecasts* with 5% - 95% (1% - 99% vertical dashed line) consistency bars. All but

Blue Dice

Jarman, Alex S. (2014) ***On the provision, reliability, and use of hurricane forecasts on various timescales*****.** **PhD thesis, LSE.**

Bröcker, J. and Smith, L.A. (2007) 'Increasing the reliability of reliability diagrams', ***Weather and Forecasting*****, 22(3): 651-661.**

April 4, 2019



FiveThirtyEight

APR. 4, 2019, AT 5:16 PM

When We Say 70 Percent, It Really Means 70 Percent

By Nate Silver

Filed under Housekeeping

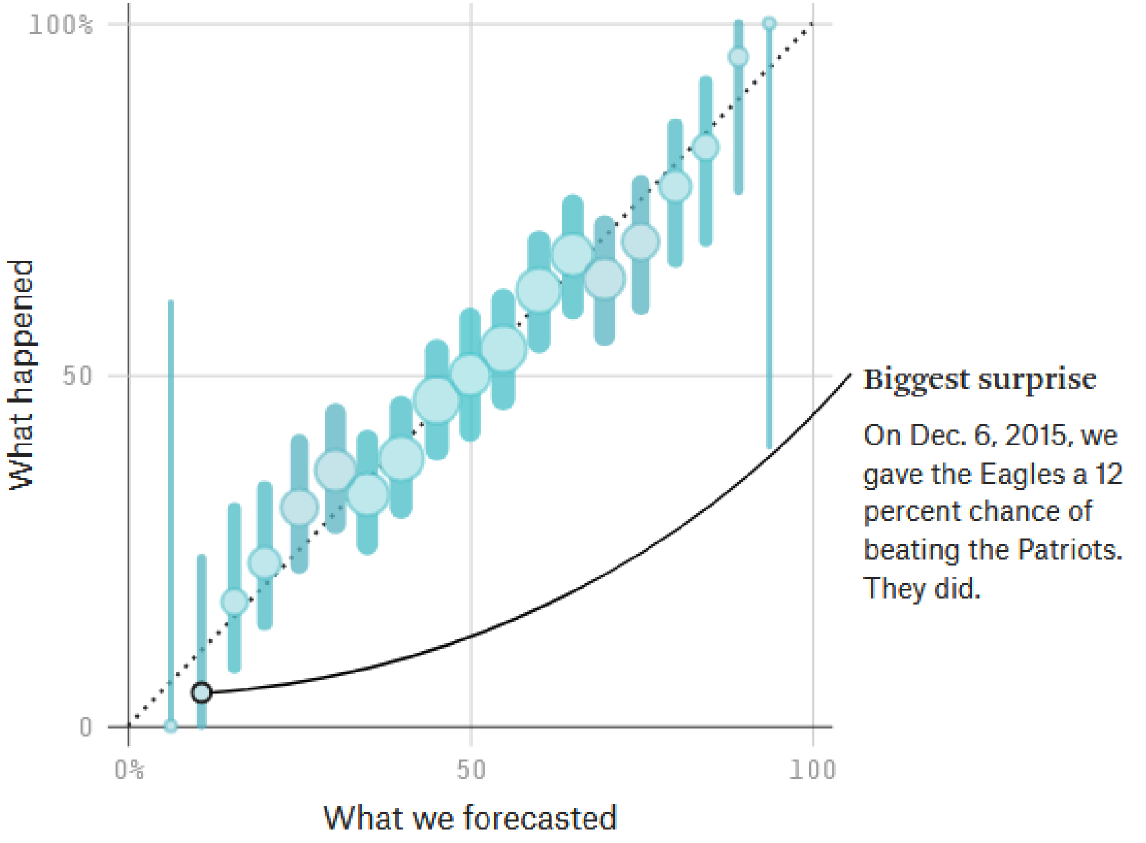
One of FiveThirtyEight's goals has always been to get people to think more carefully about probability.

Probabilistic thinking is more common in England than in the US.
There are many opportunities for outreach: the NFL and sports more generally is an excellent opportunity.

This reliability diagram is simply constructed incorrectly.

This venue offers a chance to work with 538 & "get people to think more carefully about probability."
Real people, not us.

Bröcker, J. and Smith, L.A. (2007) 'Increasing the reliability of reliability diagrams', *Weather and Forecasting*, 22(3): 651-661.



Biggest surprise

On Dec. 6, 2015, we gave the Eagles a 12 percent chance of beating the Patriots. They did.