



Project Title: ECOPOTENTIAL: IMPROVING FUTURE ECOSYSTEM BENEFITS THROUGH EARTH OBSERVATIONS

Project number: 641762

Project Acronym: ECOPOTENTIAL

Proposal full title: IMPROVING FUTURE ECOSYSTEM BENEFITS THROUGH EARTH OBSERVATIONS

Type: Research and innovation actions

Work program topics addressed: SC5-16-2014: "Making Earth Observation and Monitoring Data usable for ecosystem modelling and services"

Deliverable No: 8.3

Methodology for testing high resolution modelling

Due date of deliverable: 31/05/2018

Actual submission date: 31/05/2018

Version: v1

Main Authors: Edward Wheatcroft (LSE), Leonard A. Smith (LSE), Erica Thompson (LSE)



This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 641762



Contents

1. Executive Summary	2
2. Simulation Models in Decision Support	3
2.1. Traditional methods for use of simulation models in decision support	3
2.2. When (not) to downscale	4
2.3. Ensemble-based Insights about Extreme Event Plausibility (GLIMPSE)	5
2.4. Moving towards the vulnerability approach	5
2.4.1. A vulnerability dynamic	6
2.4.2. Example: A vulnerability approach to management of reindeer population	6
3. Data Assimilation with Imperfect Models	7
3.1. Introduction	7
3.2. Pseudo-orbit Data Assimilation Slide Set 1	7
3.3. Pseudo-orbit Data Assimilation Slide Set 2	24
4. Benchmarking Probabilistic Forecasts in Ecology	32
4.1. Climatology	33
4.2. Persistence	33
4.3. Null Models	33
4.4. Example- Benchmarking Ibex Population Forecasts	33
5. A Sanity Check For Model Selection	36
5.1. Permutation Tests	36
5.2. Sanity-Check Test	36
5.3. Ibex Population Models in Gran Paradiso National Park	37
6. Modelling the Wild Reindeer Population of Hardangervidda	38
6.1. Introduction	38
6.2. Methods	38
6.2.1. Data Sources and Preparation	38
6.2.2. Population Models and Forecasting	38
6.2.3. Testing For Density Dependence and Other Drivers of Population Change	39
6.2.4. Density Dependence	39
6.2.5. Benchmark Models	40
6.2.6. Model Selection	41
6.3. Results	41
6.3.1. Choosing a Benchmark Model	41
6.3.2. Testing For Density Dependence	41
6.3.3. Effects of Hunting	42
6.3.4. Climate Effects	42
6.3.5. Model Selection	45
7. Conclusion and Summary of Findings	49
7.1. Conclusions on Population Modelling	49
7.2. Some Reflections on Modelling	49
A. Model Selection for Correlated Time Series	53
B. Background Methodology For Population Modelling	55
B.1. Population Models	55
B.2. Model Selection Techniques	55
B.2.1. Information Criteria	56
B.2.2. Cross Validation	56
B.3. Probabilistic Forecast Evaluation	56
B.4. Assessing Model Significance	56



B.5. Ibex in Gran Paradiso	57
B.5.1. Data	57
C. Understanding Uncertainty in Environmental Modelling	58
C.1. Workshop description, aims and objectives	58
C.2. Workshop content	58
D. UUEM Lecture Slides and Transcript	63

1. Executive Summary

The aim of this deliverable is to improve the usage and interpretation of models, particularly in an ecological setting, but also in a wider context. A large part of the deliverable focuses on technical issues regarding the construction, evaluation and comparison of forecasts formed using stochastic population models. The aim of this is to review existing methodology, find areas in which there is room for improvement and suggest alternative approaches, when appropriate. It is hoped that this will lead to a better understanding of some of the technical issues behind population modelling and thus yield more confidence or caution where appropriate. For example, it is discussed how, in population modelling, it is common to propose a wide range of models and use model selection techniques to choose one or more 'best' models. In most population modelling experiments, however, it is not clear how to distinguish whether the 'best' model occurred by chance (i.e. from comparing a large number of models) or whether there is genuinely value in that model. A sanity-check is thus proposed to estimate the probability that the model selection statistic of the 'best' model could have occurred by chance.

This deliverable also aims to cast some light on the interpretation of models and model output and how these should be used in real world decision making. Whilst this is discussed in the context of ecological modelling, it is also more widely applicable. Whilst it is often acknowledged that simply choosing a 'best' model from a set of candidate models and using that to make projections into the future can be flawed, there is often a lack of ideas as to alternative, more appropriate, approaches. This deliverable aims to give insight into how one might proceed in such a situation. One approach is to choose a desired, or undesired outcome, and to find which paths might lead to such an outcome. It is hoped that these kinds of insights can lead to better understanding of how models can be interpreted and encourage model users to think more carefully about the limitations of their models.

Data assimilation is able to greatly improve the skill of forecasts formed using imperfect models, including in an ecological setting. Often, however, unrealistic assumptions are made leading to limitations in the effectiveness of data assimilation schemes. In this deliverable, a data assimilation scheme called Pseudo-orbit data assimilation (PDA) and its potential for use with ecological models is discussed. Unlike many data assimilation schemes, PDA does not assume that the model is linear and can be adapted for use with imperfect models.

This deliverable is structured as follows. In chapter 2, the question of how simulation models can be used to aid decision support in ecology is discussed. Chapter 3 discusses data assimilation. Chapters 4 to 6 focus on technical issues in population modelling. More specifically, chapter 4 deals with the issue of benchmarking, that is finding simple statistical models based only on observations. The skill of these benchmark models then acts as a baseline for the skill of a set of model based forecasts. Chapter 5 introduces a sanity-check test which, to be used alongside model selection procedures, estimates the probability that the model selection statistic of the 'best' model occurred by chance (rather than through genuine skill). The methodology in chapters 4 and 5 is demonstrated on an existing population modelling example concerning the population of ibex in Gran Paradiso National Park. Chapter 6 is a more technical version of a paper on modelling the wild reindeer population of Hardangervidda National Park written jointly with the University of Bergen and Consiglio Nazionale delle Ricerche (CNR) and to be submitted in summer 2018. Both the paper and chapter 6 use the methods introduced in chapters 4 and 5. Chapter 7 is left for conclusions and a summary of results. Appendix A discusses the issue of model selection when a set of outcomes are correlated. Appendix B describes background methodology for the population modelling of chapters 4 to 6. Appendix C summarises a workshop entitled 'Understanding Uncertainty in Environmental Modelling' (UUEM) held at LSE in Autumn 2017. The materials of the workshop are highly relevant to ECOPOTENTIAL and could easily be adapted to suit participants of the project. Appendix D is a transcript of Leonard Smith's presentation at that workshop with accompanying slides.

2. Simulation Models in Decision Support

2.1. Traditional methods for use of simulation models in decision support

Scientific simulation models are ubiquitous in evidence-informed decision making. ECOPOTENTIAL aims to make use of simulation models among other tools to support best practice for decision making about management of terrestrial and marine ecosystems in European Protected Areas. While we will consider the management of populations of reindeer and ibex explicitly here as examples, the questions discussed below arise whenever computer simulations are employed in the support of decision making.

Uncertainty Quantification (UQ) [1, 2] is an essential element in the rational application of computer simulations to decision making. The ideal aim of UQ is to provide decision-relevant probability distributions for the actual targets of interest, but a more commonly achieved goal is the quantification of the diversity of our models and their sensitivity. In the statistical community, UQ is incomplete until it includes guidance on how to apply its outputs in practice[3]; this means interpreting the model-state variables (say, modelled population levels) in terms of their real-world counter parts (actual populations). This is considered more straightforward in statistical models where the observations are taken to be the model-state variables (like the reindeer and ibex population models below) than in physics-based simulation models (like weather and climate models).

Models can support decision-making practices in many different ways. In textbook cases, reliable probability density functions (PDFs) of the target are provided and, in these cases, decision-making is straightforward, given that the desired outcome(s) is known. In practice, it is often the case that model-based PDFs do not reflect the probability of various outcomes in reality, due to known or unknown imperfection, errors or biases in the model. This may in some cases be demonstrated explicitly, where many forecast-outcome pairs are available (these cases are called “weather-like”). Even the most skilful weather forecasts typically fail (to some extent) tests of this kind of calibration, but given sufficient information it may be possible to correct for some of the effects of this inadequacy. In “climate-like” cases, where there is no large forecast-outcome archive with which to assess the quality of the probability forecasts, confidence in the relevance of the forecast probability distribution must be built on the robustness of simulation and a belief in the sufficiency of the underlying physical understanding of the system.

The native grids of global climate models are often thought too coarse to address the questions of ECOPOTENTIAL practitioners directly, so there is motivation for global model data to be “downscaled” to more relevant space and time scales. Naive downscaling approaches may result in loss of information in favour of plausible appearances, so it is critical to verify that information in the global models is preserved during this process. Where this good practice is not followed, even informative simulations in the global models will be translated into misinformative guidance at high resolution. Consistency tests which provide necessary conditions for meaningful downscaling are discussed below in Section 2.2, additional information can be found in [4].

It is possible, of course, that decision-relevant probabilities are not available in any particular case. Predictive distributions may still be informative, and knowledge of the sensitivity of the model itself remains of use in discouraging an over-precise interpretation of the simulation, but how precise an interpretation is justified? This is the question that the uncertainty guidance component of UQ addresses. It is important for ECOPOTENTIAL to provide such guidance on the appropriate use of modelled output, identifying where it can be used directly as a genuine best-estimate of real-world outcomes, and, conversely, giving more interpretation and support where the models are thought to be usefully indicative but known to have some inadequacy, error or bias. This might take the form, for instance, of the type of statement used by the Intergovernmental Panel on Climate Change when making statements about 21st century global mean temperature change based on models: they judge that a 90% confidence interval for modelled output represents only a 66% (“likely”) range of real-world outcomes. This is a subjective expert judgement presented as uncertainty guidance in a usable, quantitative form.

Alternatively, it is also possible that the available class of models is expected to be unable to make any decision-relevant quantitative insight (beyond those known before the models were run). This is not a problematic situation: in this case, the model experiments can still usefully inform scientific thought as to the nature of the dynamics of the system, the manner in which the model suggests the system works [5]. As illustrated in the reindeer case below (chapter 6), even an implausible statistical model can serve as a catalyst in identifying which additional observations should be taken or which other insights can be made. This may then help to identify improvements to the model which will develop it towards a state where the quantitative outputs could be used more directly. Relevant issues of parameter selection and data assimilation differ across these cases.

In summary, simulation models can be informative in at least four ways:

1. The provision of decision-relevant probability density functions (PDFs)
2. The provision of a general indication of the likely sensitivity of the model
3. The provision of dynamical insights into the system
4. The suggestion of which observations might distinguish between a proposed causal pathway.

2.2. When (not) to downscale

Some ecological models expect as an input detailed weather data, since this is often a key determinant of population success or failure over a season. Weather also varies a lot from year to year, varies by region or even very locally, and the statistics of weather are also changing over time (climate change).

For time series in the past, weather data may be available either in the form of observations recorded by a weather station or in the form of re-analysed gridded data which are generated by an interpolation model from observational data over the region and time period in question. Either of these can be useful in testing hypotheses about weather-related influences on population changes. A difficulty is reached at the point of extrapolating the effect of these influences forward in time in order to make projections about the relative success of populations under climatically different conditions, for example at the end of this century.

Forward projections of climate variables are provided by complex numerical global climate models (henceforth GCMs) which are run on sophisticated computers and describe the atmospheric and ocean circulation over the whole planet as well as many other earth system processes including some biology. They run at a scale of hundreds of kilometres and are thought to represent very well the processes which lead to global-scale changes in climate. The IPCC note, however, that “on regional scales, the confidence in model capability to simulate surface temperature is less than for the larger scales”, and it is generally the case that we have greater confidence in model output when it is averaged over larger spatial scales and longer time scales.

By contrast, the scales on which most decision-makers are interested in the output of complex weather and climate models are very local and often very short-term. The reindeer population in Hardangervidda does not respond to global average temperature but (perhaps) to the amount of snow lying on the ground during the leanest months, or to the favourability of weather conditions for the growth of food plants in the spring calving season. Even for global mean temperatures, GCMs give a diversity of outputs. The agreement between different state-of-the-art models is shown by the IPCC’s most recent Assessment Report in figure 9.8 ([6]), reproduced here as Figure 2.1. The graph shows considerable agreement about the magnitude of annual mean global mean temperature “anomalies” (changes with respect to the reference period of 1961-1990) but less agreement on absolute values. A disagreement in absolute value of temperature, for instance, may be unimportant for some regional studies but will become important in areas where the temperature reaches freezing (close to 0 degrees in some models but evidently up to 1.5C if anomalies are being used), or for studies where precipitation is an important factor (since the ability of the air to hold water vapour is strongly dependent on temperature).

Although global averages are the quantities we can most confidently project, there is a strong interest among decision-makers for reliable and accurate information on small local scales about a wealth of variables relevant to different kinds of decisions - information which is simply not available from a global climate model (GCM) with a grid resolution of hundreds of kilometres. Downscaling models aim to fill this gap either by generating statistical relationships between GCM and local observations based on past data (statistical downscaling), or by using a second dynamical model with finer resolution to interpolate greater detail between grid points of the GCM acting as boundary conditions (dynamical downscaling). A paper currently (May 2018) in review for Nature Climate Change [4] discusses some challenges for downscaling approaches, setting out five consistency tests which are necessary, though not sufficient, to ensure a firm practical basis for actionable outcomes.

Essentially, these consistency tests require that the outputs generated by the downscaling procedure are not overall inconsistent with the GCM outputs used to generate them. Of course, we expect that the downscaling model will introduce greater detail: for example, in mountainous areas where there is great variation in topography within the grid resolution of a GCM, we expect the downscaled output to show a much more realistic visual picture of variation which demonstrates the key topographic features. This is the aim of the downscaling approach, and within each region it may make very large corrections to the coarse GCM output without introducing any inconsistency. Physical inconsistencies can and do arise, however, at the scale of the driving GCM. For example, if a downscaling model results in 20% more precipitation within a GCM grid cell, this is water vapour being removed from the air mass which is not accounted for in the GCM and thus is available in the next grid cell to be precipitated again by the downscaling model over that area - without a consistency constraint, we may end up double-counting in this way multiple times. Regional inconsistency goes against physical conservation laws and ignores the necessity of feedbacks from one scale to the next which are normally used to maintain a physical balance. Our paper discusses these consistency requirements in greater detail and illustrates with some carefully-chosen examples in regional weather patterns due to topography, regional precipitation totals, and net radiation

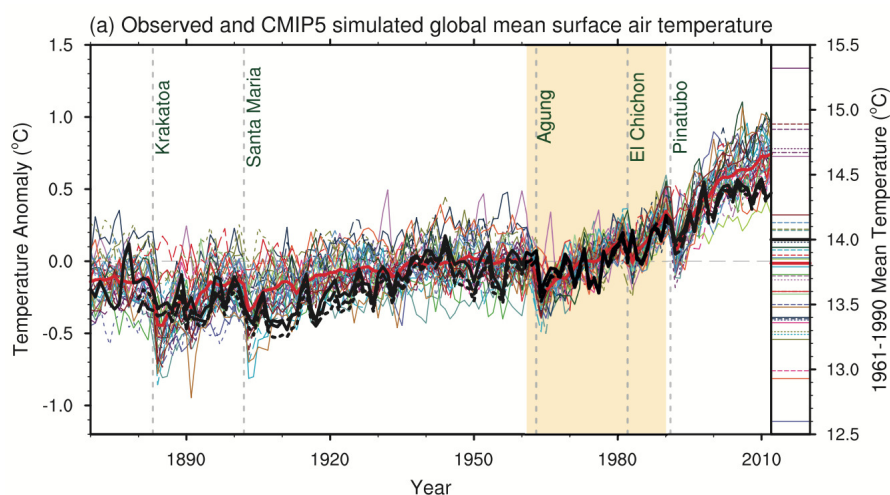


Figure 2.1: Observed and simulated time series of the anomalies in annual global mean surface temperature.

“Anomalies” are differences from the reference period (1961-1990, indicated by yellow shading) time-mean of each individual time series and the scale on the right hand side shows the actual value (degrees C) of the annual global mean surface temperature for each model over the reference period, indicating that an offset of up to 1.5C is applied. Vertical dashed grey lines represent times of major volcanic eruptions. Single simulations from CMIP5 models are shown as thin lines, the multi-model mean is a thick red line, and three alternative observational data sources (HadCRUT4, GISTEMP and MLOST) are in black. All simulations use specified historical forcings up to and including 2005 and use RCP4.5 thereafter. From the IPCC Fifth Assessment Report, 2013, Chapter 9: Evaluation of Climate Models, figure 9.8.

balance over an area. The examples demonstrate the ways in which a downscaling technique might fail our consistency tests and the implications of doing so, pointing the way for careful but not overly onerous evaluation of new methods.

Where ECOPOTENTIAL research projects use the outputs of such downscaling models, therefore, we recommend that these consistency tests are applied carefully to ensure no immediate red flags are raised about the suitability of the information for decision support.

It is worth noting that downscaled data will almost universally appear more visually compelling regardless of the quantitative accuracy of the information presented: it is much easier to believe a colourful, detailed map on which you can identify familiar landscape features than one which consists of a patchwork of 100-km grid squares. Nevertheless, we must remember that precision is not the same as accuracy - it is possible that the downscaling method offers no added value apart from visual appeal, and it may even be a quantitatively worse source of information. Methods such as those presented in this document can then be used, where suitable data are available, to evaluate the quantitative accuracy of some forecasts.

A paper entitled “Considerations for Users and Providers of Downscaled Climate projections”(Thompson, 2018), which covers similar topics to those described above, has been submitted to Wiley Interdisciplinary Reviews:Climate Change.

2.3. Ensemble-based Insights about Extreme Event Plausibility (GLIMPSE)

Forewarning of extreme conditions (be they weather events or harsh seasons) can prove of great use in the vulnerability approach. LSE has been laying the foundations for a method which targets forewarning rather than forecasting; the approach does not attempt to provide the probability of an extreme event, but rather to raise a warning when that event becomes plausible. The plausible event can then be tracked as time passes, and mitigatory action contemplated. This work will be reported under ECOPOTENTIAL Work Package 2 but is closely related to the modelling work of Work Package 8.

2.4. Moving towards the vulnerability approach

The discussion above assumes a predict-optimize-act dynamic. This is, of course, not the only manner of using information to support decisions and in the absence of decision-relevant PDFs, ECOPOTENTIAL stresses the fact that alternatives may be vastly preferable. Here, we outline a version of the “vulnerability dynamic” which might offer superior decision

support in many of the contexts of interest to ECO POTENTIAL partners/practitioners.

2.4.1. A vulnerability dynamic

A vulnerability dynamic considers the desired outcome(s), and then asks what events would disrupt, prevent the outcome, or make achieving it significantly less probable. The outcome may be a thriving population of reindeer in Hardangervidda, the smooth operation of a coastal Nuclear Power Station, maintenance of manageable levels of demand for health services, happy surfers (no jellyfish), and so on (defined by the practitioner). Disruptive events in wildlife management would include an overly-large cull (too much hunting/trapping/fishing), starvation in the winter, dehydration/heat-stroke death in the summer, low birth rates, and so on (an initial list would be defined by the practitioner, and more potential disruptive events may become apparent in the modelling process).

Then, we select a disruptive event. Identify:

1. its immediate impact on the desired outcome(s);
2. susceptibility of the outcome(s) to a single event;
3. options for mitigating its impact (if any);
4. efficacy of advance warning of impact (required lead times to take mitigation action);
5. robustness (recovery time) of the target;
6. options for increasing speed of recovery.

The aim is to consider the vulnerability of the desired outcome: what places it at which levels of risk? If recovery is possible, how costly would aiding recovery be relative to mitigation? What management strategies lead to survival/the desired outcome.

This allows something of a rank ordering to disruptive events, from most deadly in terms of impact on the desired outcome) to less so. Select the most deadly event. Only at this point is information from model simulations considered. What is the background rate for this event, and how well is it known? How high is the fidelity of current models regarding this event? Would advance (“weather-like” probabilistic) warning be of use? What rate of this event would put achieving the outcome in doubt? Are there general management practices that would increase the outcome’s likely survivability of the event? How might changes in frequency of the event arise? Are there first-principles limits?

One can *then* ask if it would be worth attempting to simulate the event. Next, we repeat the process for the second most deadly event.

2.4.2. Example: A vulnerability approach to management of reindeer population

Ecological management decisions can only be made where there is a preferred outcome. If the preferred outcome is not defined (such as for management of the reindeer population in Hardangervidda), then no amount of information about the ecological dynamics of the population will determine an “optimal” strategy¹.

Where a preferred outcome is defined, for example

- maintenance of a population above a genetically viable level;
- maintenance of a population below an environmentally destructive level;
- maintenance of a population that can support a certain level of harvesting or recreational hunting each year;
- or some combination of these and other factors, then information about the dynamics of the system will be helpful, characterised as “If we do [action X] then [outcome Y] will result”.

We can then look at all of the outcomes of the range of potential actions which are open to us, and choose the one we prefer most - in doing so, we could also allow for some uncertainty in these outcomes by describing them probabilistically and evaluating the merits of the choice with respect to some more mathematically-precise utility function including our attitude towards risk. For example, perhaps in pursuit of other goals we are willing to accept a small probability of the population becoming too high in any given year, but we will accept no risk at all of extinction since this would be a catastrophic and irreversible event.

If ecological population models were perfectly accurate then they would be a direct input to the decision-making process in this way. As described in this document, however, there are many sources of uncertainty which cannot be fully captured within the modelling process: selection of parameters and indeed selection of models are always subject to a limitation that the true dynamics are not actually described by anything within the candidate set of models. Data assimilation and all statistical methods are limited by the existence of a sufficiently large set of sufficiently relevant data. When there are only 21 data points from which to construct a model, it is inevitable that there will be great uncertainty in the identification of any significant dynamical relationships.

¹unless we have implicitly defined a non-zero population as the goal, and by some fluke it happens to be the case that there is only one choice of management strategy that leads to anything other than a non-zero population.

3. Data Assimilation with Imperfect Models

3.1. Introduction

Data assimilation is often viewed only as the task of moving from observations of the real-world to initial conditions that are viable in the state space of a model. Most operational methods make unrealistic assumptions regarding the perfection of the model, the stochastic nature of the dynamics, or both.

LSE CATS has spearheaded the development of a new approach, Pseudo-Orbit Data Assimilation (PDA), based on the previous work of Judd and Smith [7, 8]. The talk transcript that follows is taken from several seminars. The slides reflect the astronomy colloquium at the University of Durham in late 2017. Since 2016, PDA has been developed to the point that it is straightforward to apply to large simulation models which have an adjoint, and is also applicable to models with no adjoint (this was previously done by Judd. More recently, working under ECOPOTENTIAL and jointly with the University of Chicago, PDA was implemented for the NCAR BGrid atmospheric model. We have not yet had the opportunity to implement it with a European model, but remain interested in doing so. The advantage of PDA is that it opens many many doors for working with real-world (and therefore imperfect) models.

The slides give a clear schematic vision of the concepts underlying the algorithm. We stress that data assimilation is much more than merely the task of moving from observations of the real-world to initial conditions that are viable in the state space of a model. PDA, along with a simple method for making pseudo-observations given the model state, allow us to move between different models in a natural way that require very little additional coding. This opens a wealth of new directions given a collections of imperfect models, as well as new ways to use the models we have, targeting vulnerabilities in an attempt to gain early warning of particular phenomena rather than probability distributions which are unlikely to be well-calibrated. Work on this application continues.

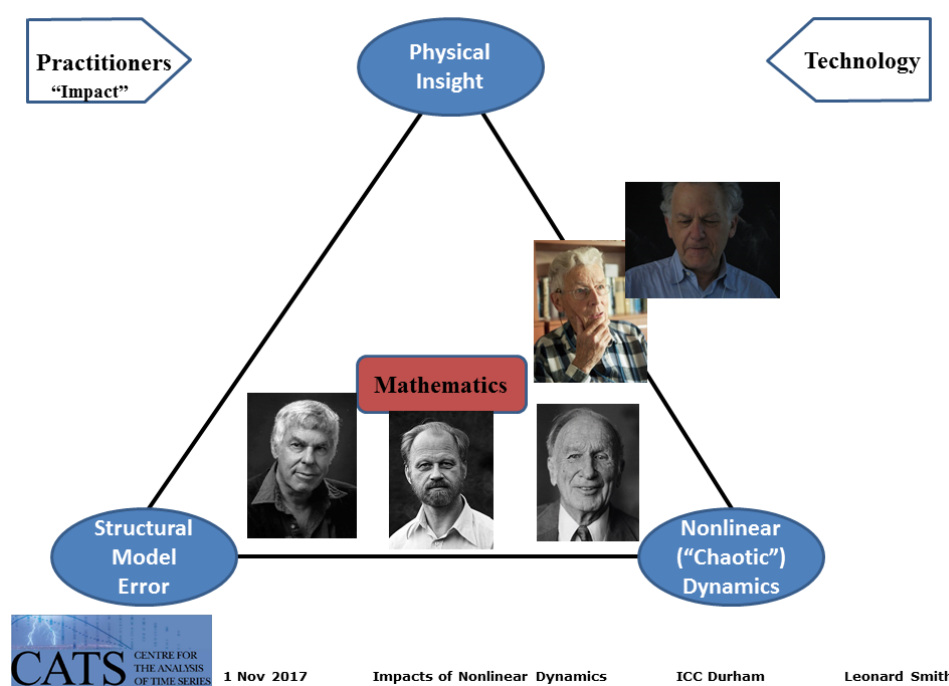
3.2. Pseudo-orbit Data Assimilation Slide Set 1

Pseudo Orbit Data Assimilation Presentation from Durham Astronomy Colloquium 2017

Leonard A Smith

Note the transcript below was taken from a recording and may contain inaccuracy, flaws, and mistakes of speech.

So now we see how physical insight, structural model error and nonlinear chaotic dynamics intervene, and in particular we see how the work of Moser and Smale actually put a limit on what we can expect when we look at the sort of systems that Lorenz and Hénon and Ed Spiegel were looking at, and that we still look at today. The question is: how do we move on? How do we make progress? And the answer, one answer, is to look at different ways of data assimilation; methods of data assimilation which actually allow for model error - in fact were built on the assumption that there is model error - and this model error does not have a simple statistical description.



It is interesting to note that people talk a lot about moving to stochastic models as if it is a way of getting around model error. Tim Palmer does this a lot. I think

there are other aspects in which Tim's ideas on stochastic physics, of putting stochastic terms into the equations used to model deterministic systems, are very good ideas. It could be much better to put a realisation of a process into a nonlinear model than a constant that is the average of that process, even if the realisation has nothing to do with what is really happening at small scales. That said, the idea in general that introducing stochasticity takes away model error is misdirected. If you put stochasticity into the model, if you've taken a deterministic model with model error and (the model error is well described by a stochastic term) then you've created a stochastic model with model error, different model error but model error none the less. And in terms of performance you actually reduced the model error. If it doesn't, if the model error has properties which are not reflected in the stochastic forcing you put in then it is not even going to reduce the impact of (the new, different) model error. It will just give you model error of a different flavour.

We want to allow for that kind of problem and that is what Pseudo Orbit Data Assimilation does (PDA). This is work both with Judd and with Du, and recently we have actually started trying to apply this to the NCAR BGrid model. Kevin Judd. With Kevin we have already applied it to the Navy's NOGAPS model, as reported here. So it actually works on large nonlinear operational systems!

Finding reference trajectory \mathbf{u}_t : model state at time $t \in \mathbb{R}^m$
 $\mathbf{i}\mathbf{u}$: point in sequence space $\mathbb{R}^{m \times n}$
 \mathbf{u} : \mathbf{u} at GD algorithmic-time i

$${}^0\mathbf{u} = \{S_{-p}, \dots, S_0\}$$

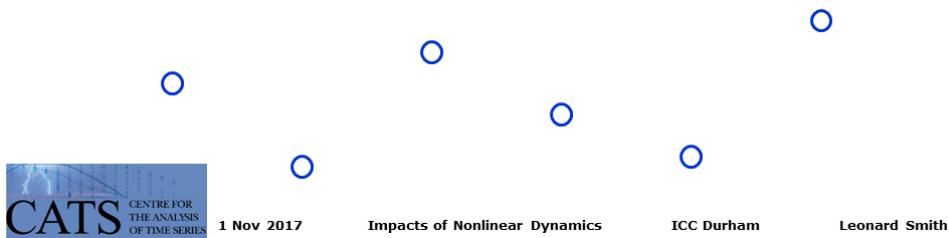
\mathbf{u} itself is a pseudo-orbit

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains any series of n model states.

Define the mismatch error cost function:

$$C_{GD}(\mathbf{u}) = \sum_{t=-n+1}^0 |F(\mathbf{u}_t) - \mathbf{u}_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.



Finding reference trajectory \mathbf{u}_t : model state at time $t \in \mathbb{R}^m$
 $\mathbf{i}\mathbf{u}$: point in sequence space $\mathbb{R}^{m \times n}$
 \mathbf{u} : \mathbf{u} at GD algorithmic-time i

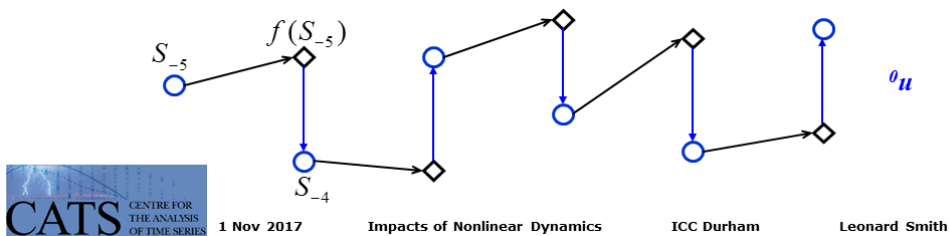
$${}^0\mathbf{u} = \{S_{-p}, \dots, S_0\}$$

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains **any** series of n model states.

Define the mismatch error cost function:

$$C_{GD}(\mathbf{u}) = \sum_{t=-n+1}^0 |F(\mathbf{u}_t) - \mathbf{u}_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.



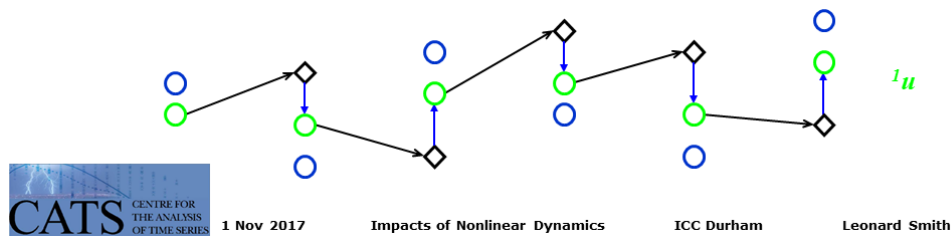
Finding reference trajectory

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains any series of n model states.

Define the mismatch error cost function:

$$C_{GD}(\mathbf{u}) = \sum_{t=-n+1}^0 |F(\mathbf{u}_t) - \mathbf{u}_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.



So what do we want to do? We want to find a reference trajectory. We want to find a place to start our model. This is one of the things that data assimilation does. It starts with observations of the real world and it somehow casts those into model land and gets us good initial conditions in model land. So let us suppose we already have some cheap form of estimate of the state of the model state at time t . This is in some high dimensional space \mathbb{R}^m . We are going to actually make that space higher.

We are going to go to a sequence space which takes a series of model states, n of them, and strings them together into one large vector. Make sense? So now we are in an m by n dimensional space and we are also going to use mu just to indicate the algorithmic time of the data assimilation algorithm. So what we really want is vector u , here, which is a series of states. That is a pseudo-orbit, even if we just put the observations in, the observations cast into model states, we will have a pseudo-orbit.

Now that pseudo-orbit is almost certainly not a trajectory. If we take one time slice and move it to the next time slice there will be some mismatch. We are going to call that 'mismatch error' where we basically find our cost function to be that mismatch error here. So what is this? We are looking at f at u of t minus u of t plus one. That is the image the current state minus the next state. So if this is a trajectory then this term in the cost function will be zero and then we are going to sum it over all points in the assimilation window, but that is just the difference between two points in the sequence space. When this goes to zero we have a trajectory and all trajectories have this equals zero.

The means that there is a manifold of points with zero mismatch error in the sequence space which corresponds to every trajectory. So we have a necessary and sufficient condition for having a trajectory. It is on this lower dimensional manifold. The vast majority of states, of course, aren't anywhere close to being the image of two randomly chosen states and we actually have a string of n states in a row which allows us to look over a very long window. What does that look like?

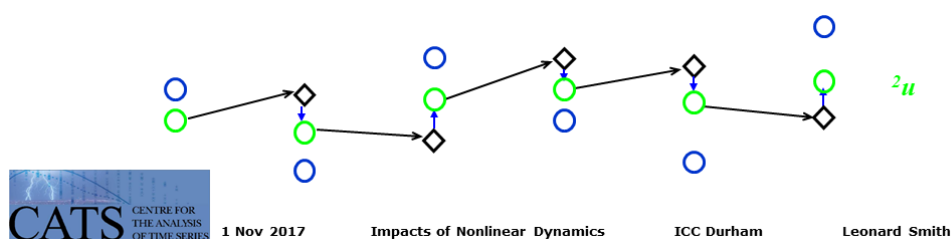
Finding reference trajectory

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains any series of n model states.

Define the mismatch error cost function:

$$C_{GD}(\mathbf{u}) = \sum_{t=-n+1}^0 |F(\mathbf{u}_t) - \mathbf{u}_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.



Well, look at the gradient descent algorithm. These blue dots, there they are, those correspond to different observations and there is some circle around them which is the observational noise. We look at s minus five. We see where it goes. That gives us the image of s minus five, and we see how far it is, the blue arrow, from s minus four. So the black arrows take us from one state to the image of that state - the diamond - and then the blue arrows show us the distance between the image of the state and the next state. That is the mismatch error.

If you look at the next sequence - there we go, we update it - and we get the green arrows. So what we have done now is we have taken a step or two in gradient descent space. What that has allowed us to do is get something with smaller mismatch errors. The question is, if we keep doing this like this it keeps shrinking until, well if the model is perfect, we will actually get a trajectory.

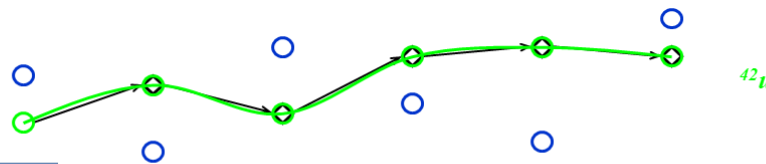
Finding reference trajectory

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains any series of n model states.

Define the mismatch error cost function:

$$C_{GD}(\mathbf{u}) = \sum_{t=-n+1}^0 |F(\mathbf{u}_t) - \mathbf{u}_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.

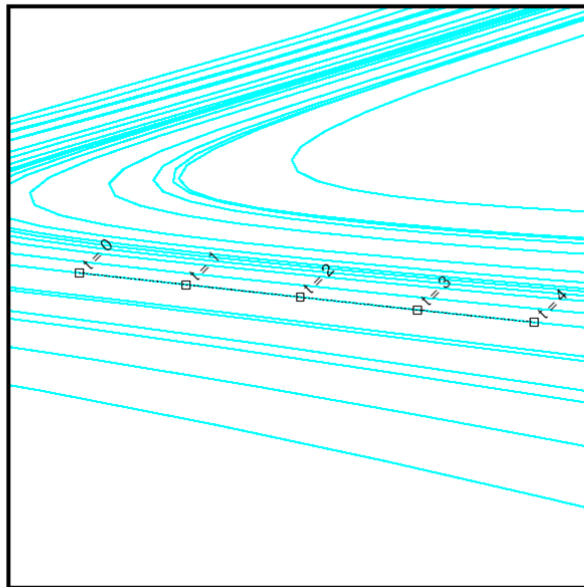


Of course, if the model is imperfect we may prefer a p-orbit to a trajectory!

In this case after forty two steps we have a trajectory of the system, we are on the manifold. Now if the model isn't perfect, of course, we don't want to go that far. We actually may prefer to have a pure bit and by looking at the residuals, the directions of those mismatch errors at the end, each component will still have a little mismatch, we can often tell what is wrong with the model.

Here is an example with Lorenz. Suppose we have five observations - t equals zero to four. Those observations don't actually lie on the attractor. Each one is sort of near where the true state is.

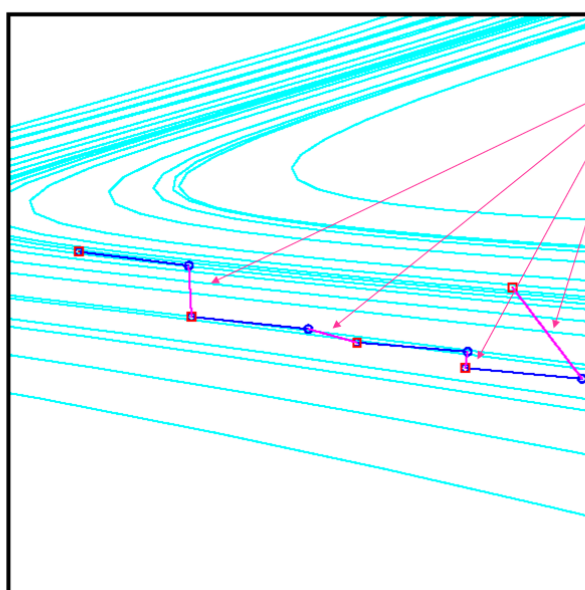
THE GEOMETRY OF MODEL ERROR



Here is a trajectory
segment of Lorenz 63

We now look forward making those one step forecasts, the blue lines. We look at the mismatch errors in, this case they are sort of purple/magenta, and the sequence - this is actually the data - the sequence pulls those back in. There you go. So we get convergence towards trajectory. We get very close. We just take the initial condition and run it forward, right?

THE GEOMETRY OF MODEL ERROR

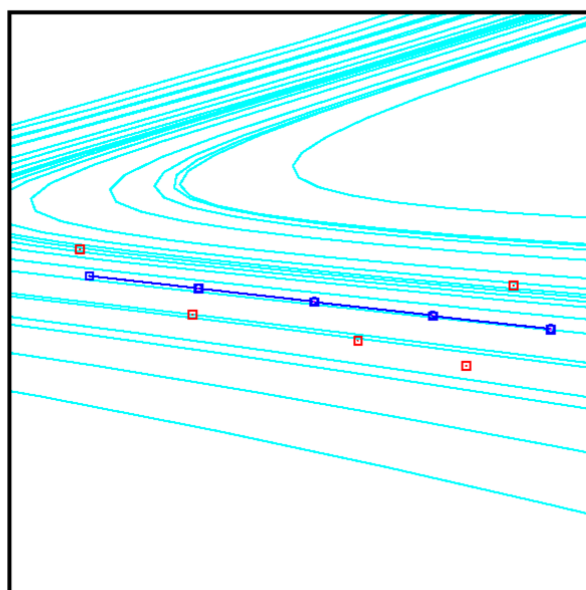


The aim is to minimize the mismatches simultaneously.

Using, say, simple gradient decent, in the 14-d sequence space, towards a global minima on the trajectory manifold.

After using the observations to define the starting point, they are ignored during the (initial) decent.

THE GEOMETRY OF MODEL ERROR

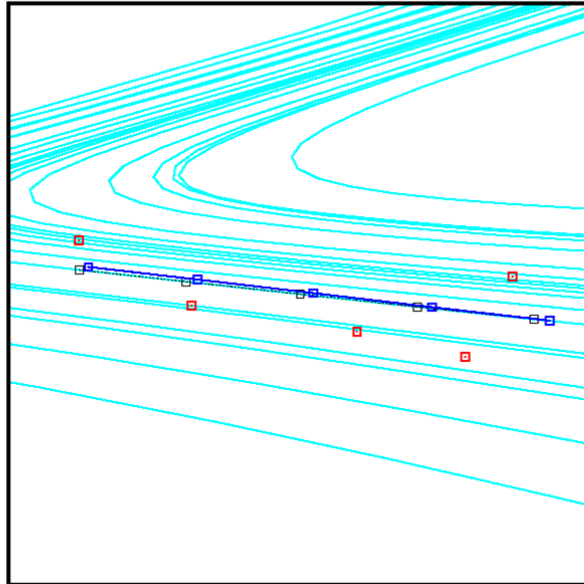


Convergence toward a trajectory.

Once very close, the trajectory passing through any point on the psuedo-orbit can be used/contrasted with other trajectories.

We can actually get trajectories or pseudo-orbits off the end. We are very close to truth, truth is the dotted line, but it is not the truth. It is instead indistinguishable states of truth. In fact if we have a decent set of indistinguishable states the probability of the truth, given this trajectory, is in fact the probability of this trajectory given truth.

THE GEOMETRY OF MODEL ERROR

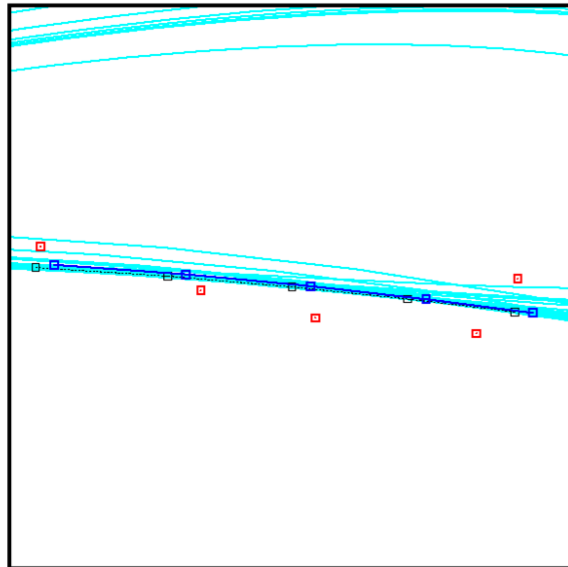


Near Truth, but not
Truth

We turn to look sideways and what we see is it actually lies, both the trajectory and the true trajectory, lie on the attractor whereas the observations simply don't. So what happens if we get an observation which is far from the attractor like this one? We kill that one.

We get this observation. Now if you are thinking about route mean square forecast errors it is going to try and minimise that giant deviation. It is going to pull everything off. If you look at a common filter it pulls off the entire ensemble. But what happens is it sees - PDA sees - that this observation, this initial condition, this point in the sequence space, is far from the trajectory manifold.

THE GEOMETRY OF MODEL ERROR



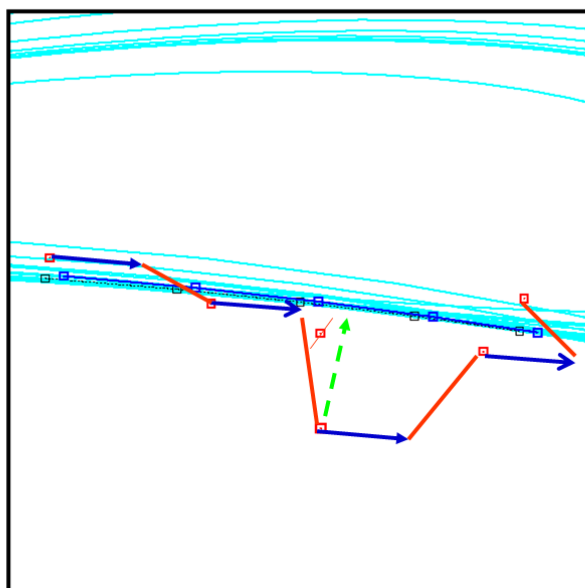
The trajectory is near
the **natural manifold**;
the obs are not!

(Near defined rather
poorly using the noise
model!)

The trajectory is also
near to (but different
from) the segment of
truth that generated the
obs.

Its image is far from the model trajectory from the manifold and the next point is actually closer.

THE GEOMETRY OF MODEL



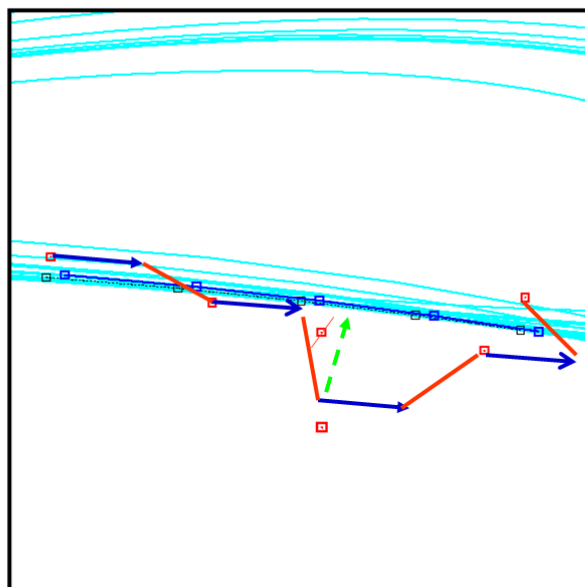
Suppose the observation at $t=3$ had been significantly in error.

The shadowing filter can recover using observations from $t=4$ and beyond, in a manner that sequential filters cannot.

In the shadowing filter, the mismatch at $t=3$ and $t=4$ is decreased by bringing the estimated

Sequential filters do not have access to this multi-step information.

THE GEOMETRY OF MODEL



Suppose the observation at $t=3$ had been significantly in error.

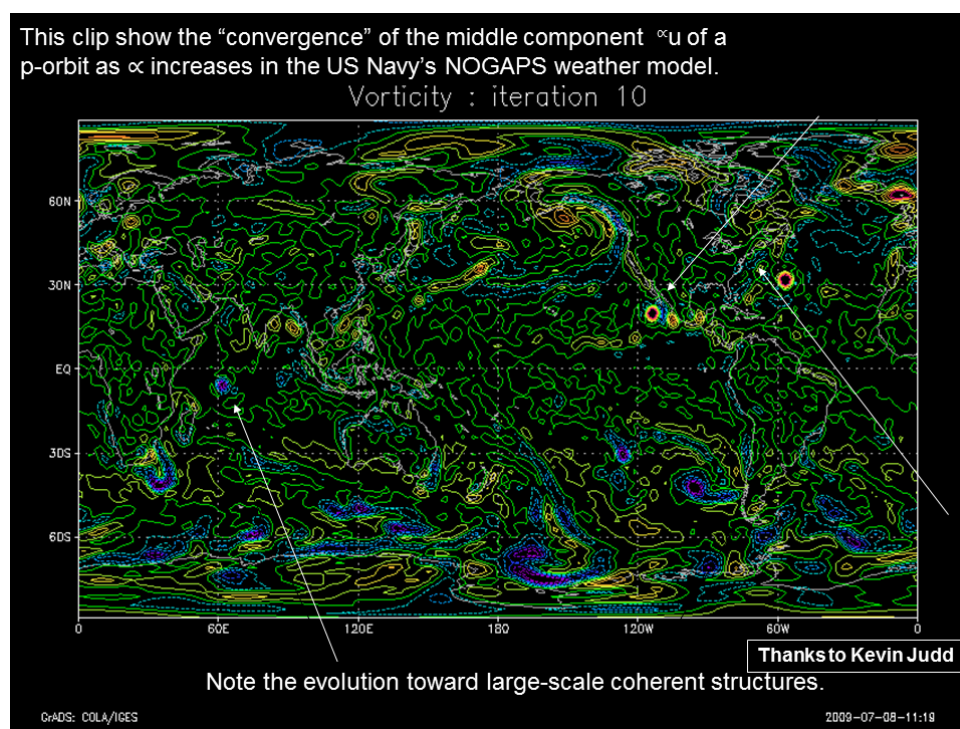
The PDA can recover using observations from $t=4$ and beyond, in a manner that sequential filters cannot.

The mismatch at $t=3$ and at $t=4$ are **both** decreased by bringing the estimated state at $t=3$ back toward the model manifold

Sequential filters do not have access to this multi-step information.

So what PDA will do - this is a schematic illustration, not a real value - is it actually brings that point back into line. Sequential filters just don't have access to this multistep information. Variational assimilation has a tendency... requires short windows to avoid local minima. Weekly constrained variational assimilation requires you actually know the noise function, and it usually assumes it is white and unrealistic. That is another paper with Tim and David Orrell. And we basically pull that point, here we can actually pull the bad point, back into the trajectory.

Next, just to show you a clip that Kevin made using NOGAPS, this is just to try and prove /demonstrate, prove by demonstration, that the system really works with very high dimensional points and here is a couple...



Operational Mismatch: Models in a 10^6 D State Space

K Judd, CA Reynolds, LA Smith & TE Rosmond (2008)
[The Geometry of Model Error](#). *Journal of Atmospheric Sciences* 65 (6), 1749-1772

D Orrell, LA Smith, T Palmer & J Barkmeijer (2001) [Model Error in Weather Forecasting](#), *Nonlinear Processes in Geophysics* 8: 357

**But first, an illustration using Newton's Laws of gravitation
and the effect of Jupiter on the inner planet**



1 Nov 2017

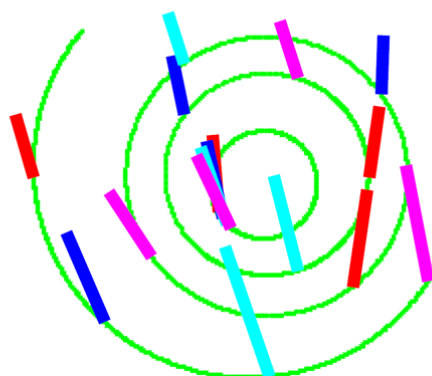
Impacts of Nonlinear Dynamics

ICC Durham

Leonard Smith

There are a couple of papers here, these are references to papers with Kevin, on The Geometry of Model Error which has the NOGAPS results and the David Orrell paper, which actually shows that we have this kind of sustained error in time in weather forecasting as well. Ok, so I want to use this to illustrate Newton's Laws of Gravitation to show how I would expect it to work in the case of actually looking at the effects of Jupiter on the inner planet or going back and sort of acting as if we are going to discover Uranus again.

Operational Mismatch: Newton



System:
Newton's Laws
Five Planets

Model:
Newton's Laws
Four Planets



1 Nov 2017

Impacts of Nonlinear Dynamics

ICC Durham

Leonard Smith



3.3. Pseudo-orbit Data Assimilation Slide Set 2

Finding reference trajectory

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains any series of n model states.

Define the mismatch error cost function:

$$C_{GD}(u) = \sum_{t=1}^{n-1} |F(u_t) - u_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.

1 Nov 2017 ICC Durham Leonard Smith

Finding reference trajectory

Given a sequence of n observations of m dimension system, we define a sequence space a $m \times n$ dimensional space, which contains any series of n model states.

Define the mismatch error cost function:

$$C_{GD}(u) = \sum_{t=1}^{n-1} |F(u_t) - u_{t+1}|^2$$

Applying a Gradient Descent algorithm, starting at the observations and evolving so as to minimise the cost function.

1 Nov 2017 ICC Durham Leonard Smith

GD course, if the model is imperfect we may prefer a p-orbit to a trajectory!

Here is a trajectory segment of Lorenz 63

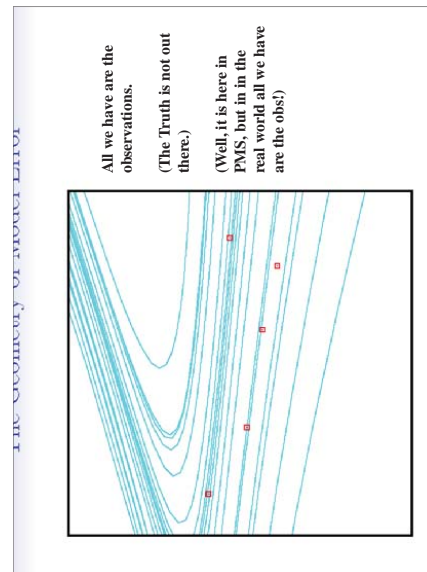
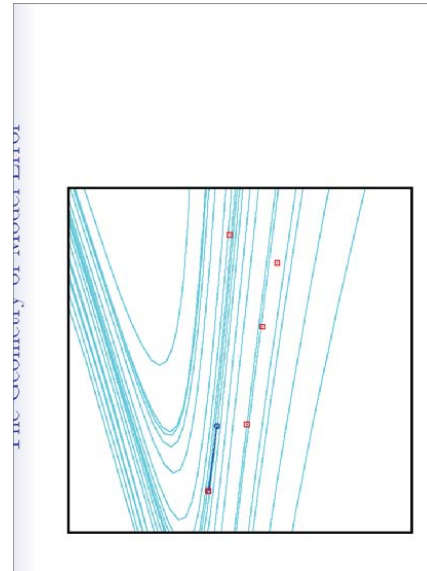
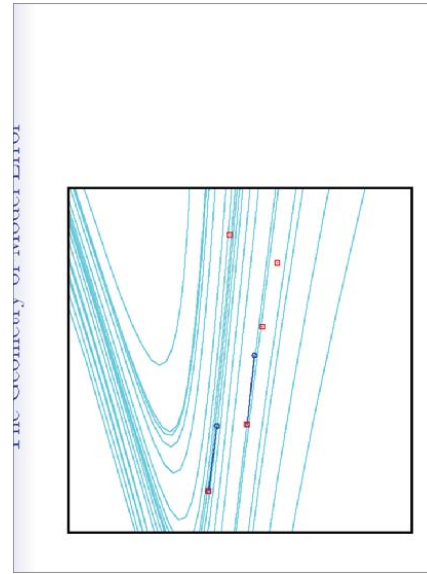
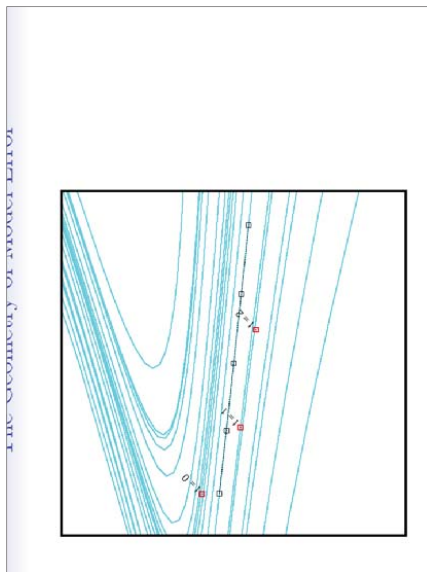
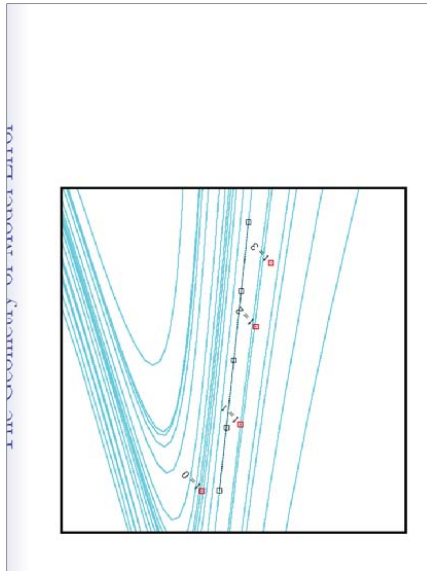
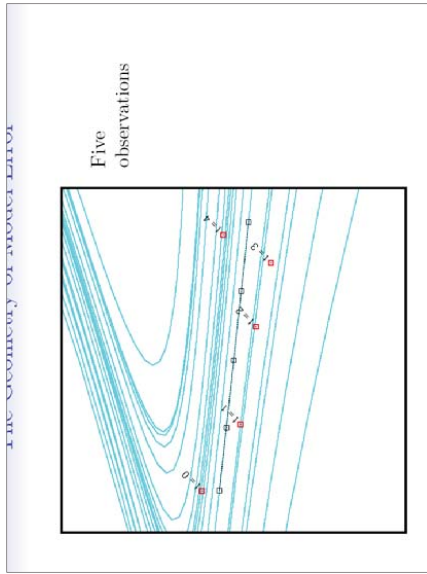
THE COMPLEXITY OF MODEL ERROR

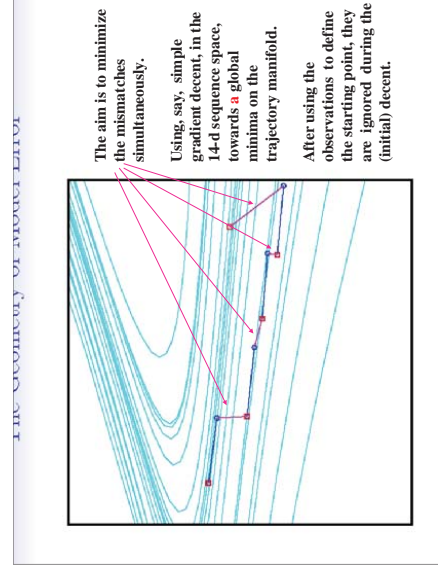
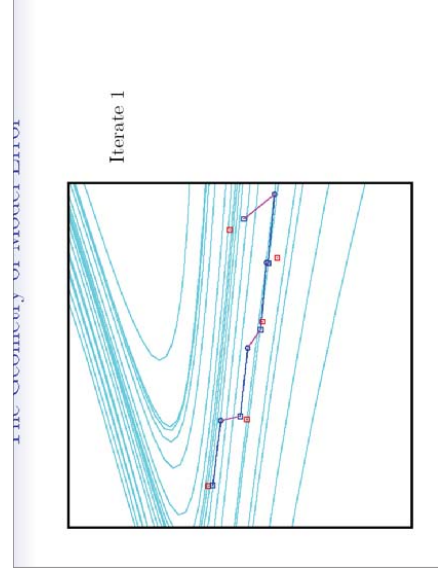
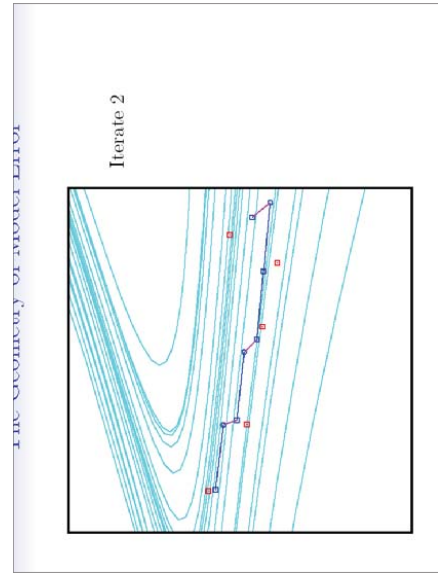
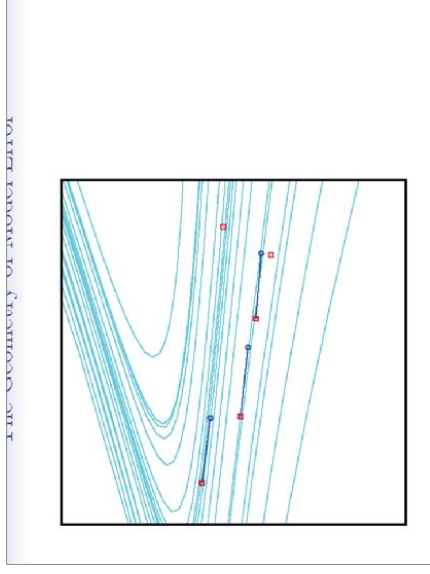
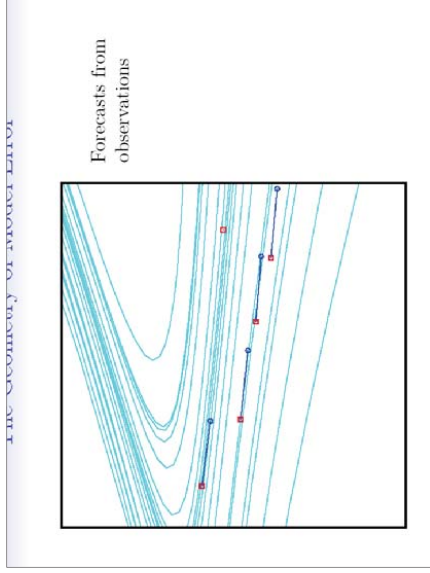
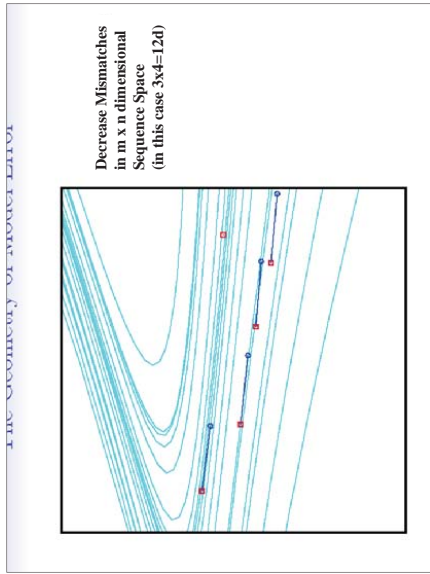
Making observations

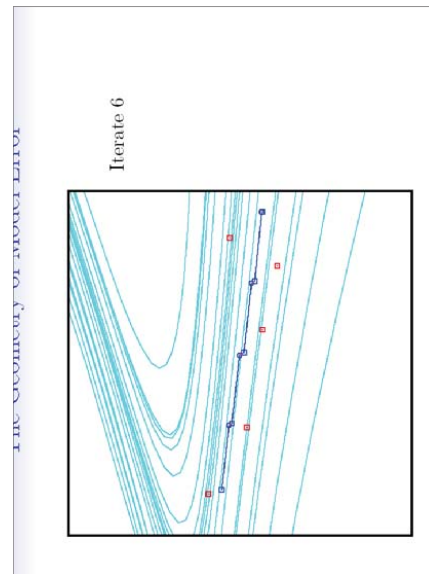
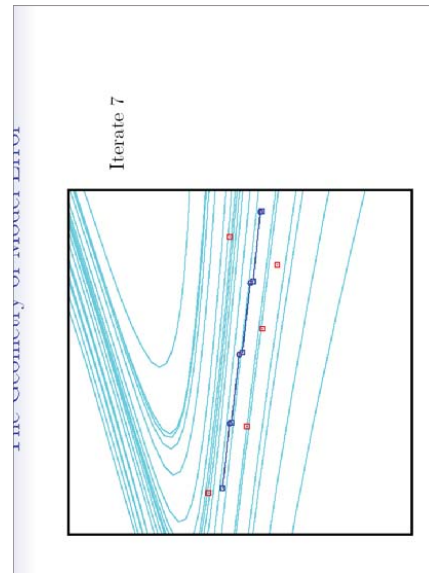
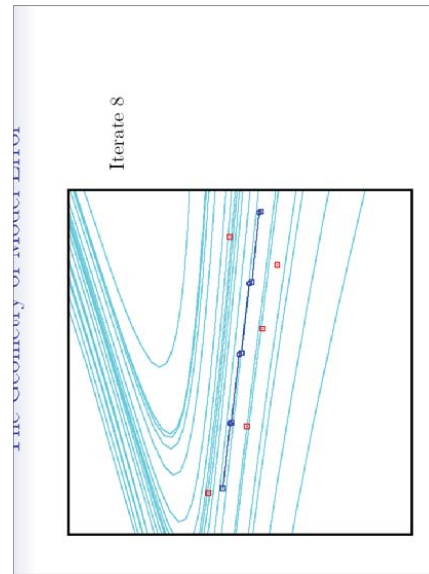
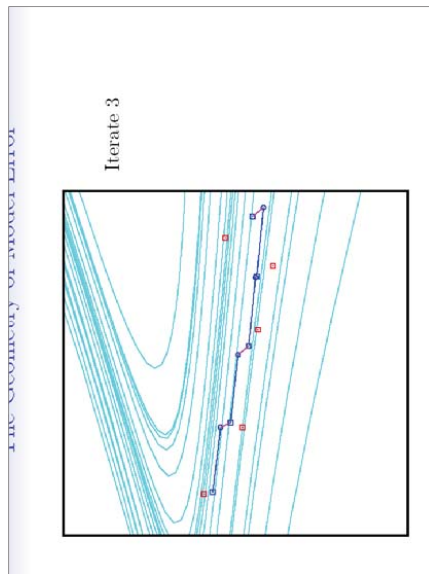
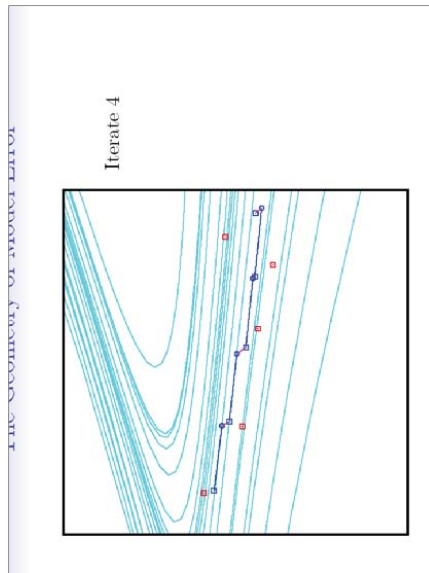
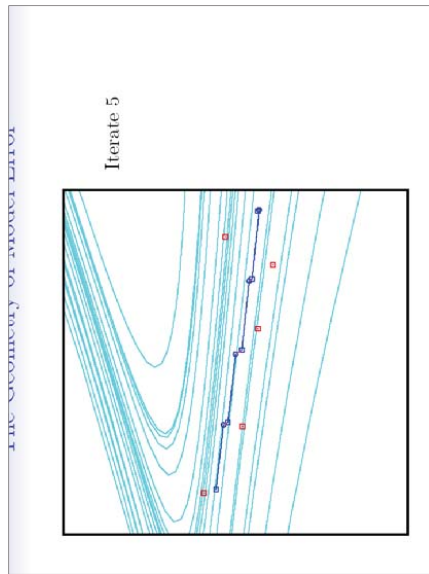
THE COMPLEXITY OF MODEL ERROR

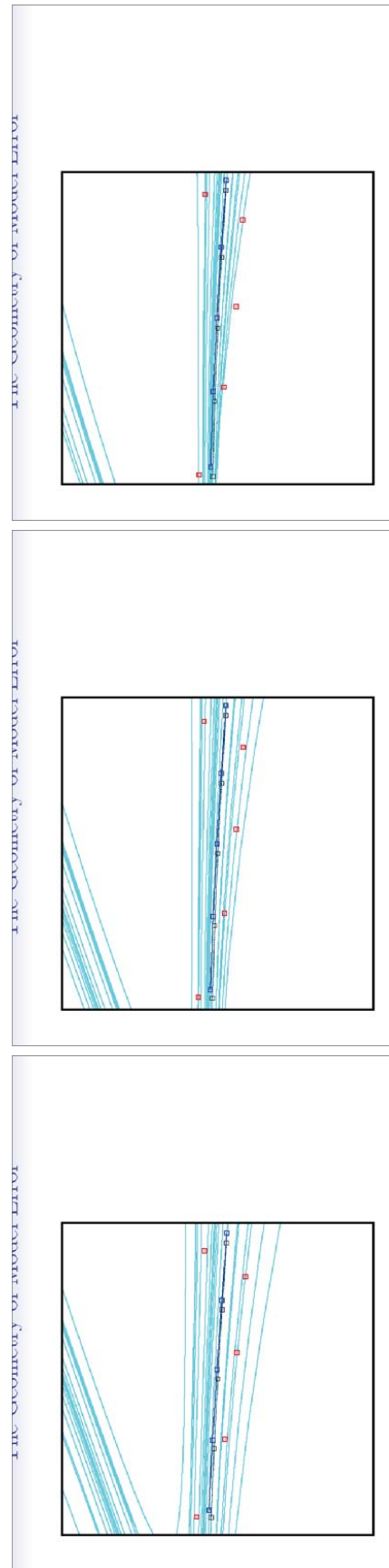
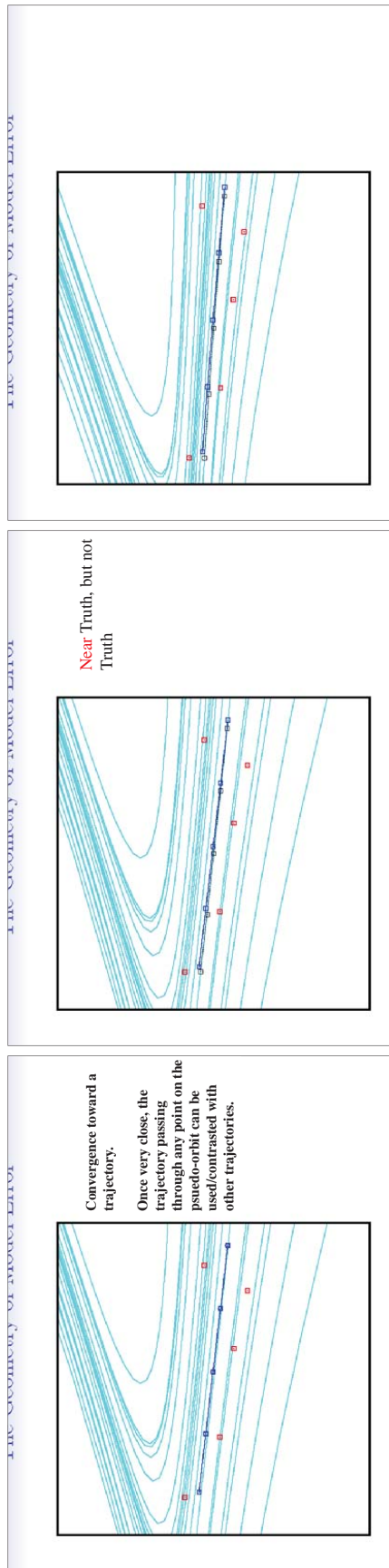
THE COMPLEXITY OF MODEL ERROR

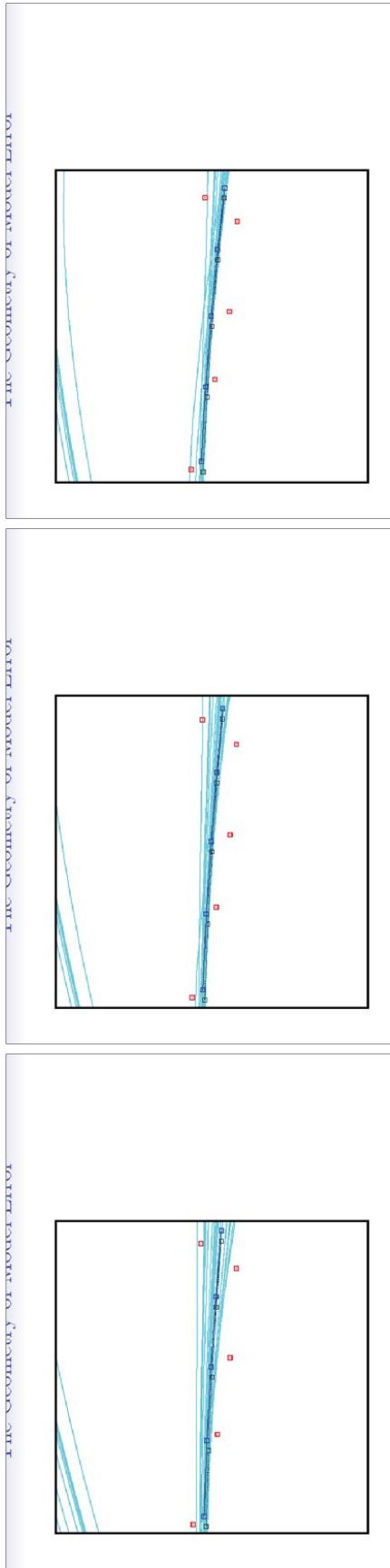
THE COMPLEXITY OF MODEL ERROR











The figure consists of three sequential plots, each labeled 'THE OCCURRENCE OF AVERAGE STATE' at the top. Each plot shows a bundle of blue lines representing a trajectory and several red squares representing observations. The trajectory is a narrow band of lines that slightly curves to the right. The observations are scattered points that generally follow the path of the trajectory.

The trajectory is near the natural manifold; the obs are not!
 (Near defined rather poorly using the noise model.)
 The trajectory is also **near to** (but different from) the segment of truth that generated the obs.

This is achieved by paying more attention to the dynamics over the window.
 Statistical properties relating the trajectory to the observations are secondary.
 This proves **remarkably robust** either:
 - when the model is perfect
 - in high-dimensional spaces

Suppose the observation at $t=3$ had been significantly in error. The shadowing filter can recover using observations from $t=4$ and beyond, in a manner that sequential filters cannot.
 In the shadowing filter, the mismatch at $t=3$ and $t=4$ is decreased by bringing the estimated state at $t=3$ back toward the model manifold

THE COHERENCY OF MODEL

Suppose the observation at $t=3$ had been significantly in error. The shadowing filter can recover using observations from $t=4$ and beyond, in a manner that sequential filters cannot.

In the shadowing filter, the mismatch at $t=3$ and $t=4$ is decreased by bringing the estimated state at $t=3$ back toward the model manifold. Sequential filters do not have access to this multi-step information.

THE COHERENCY OF MODEL

Suppose the observation at $t=3$ had been significantly in error. The PDA can recover using observations from $t=4$ and beyond, in a manner that sequential filters cannot.

The mismatch at $t=3$ and $t=4$ are both decreased by bringing the estimated state at $t=3$ back toward the model manifold. Sequential filters do not have access to this multi-step information.

THE COHERENCY OF MODEL

Suppose the observation at $t=3$ had been significantly in error. The PDA can recover using observations from $t=4$ and beyond, in a manner that sequential filters cannot.

The mismatch at $t=3$ and $t=4$ are both decreased by bringing the estimated state at $t=3$ back toward the model manifold. Sequential filters do not have access to this multi-step information.

THE COHERENCY OF MODEL

Given that we can find one such trajectory near the obs, we can create an ensemble from the set of indistinguishable states of that (and similar) trajectories, and then draw from that set conditioned on how well each member compares with the observations.

David A. Smith, Physics D, University of California, San Diego, La Jolla, CA 92037, United States of America, 2004. Da. S. Smith, JAS, 2014.

The aim of data assimilation in PMS is an accountable probability forecast.

Moving Forward: Operational Mismatch

CATS
1 Nov 2017
Impacts of Nonlinear Dynamics
ICC Durham
Leonard Smith

Thanks to Kevin Judd

This clip show the "convergence" of the middle component ψ_u of a p -orbit as α increases in the US Navy's NOGAPS weather model.

Vorticity : Iteration 10

Note the evolution toward large-scale coherent structures.

GEOS SOLUTIONS
300-25-01-11-18

4. Benchmarking Probabilistic Forecasts in Ecology

It is common in ecology to consider multiple candidate models to describe the same process. Different models may consist of different combinations of predictor variables in a generalised linear model, for example. Alternatively, each model may be structurally different, treating important processes in a slightly different way, perhaps. Commonly, candidate models are compared using model selection techniques such as information criteria [9, 10] and cross-validation [11] (both described in the appendix). Whilst these approaches attempt to measure the relative performance of a set of models, in general, they provide little evidence regarding their absolute performance, i.e. whether they provide a useful fit to the data at all. After all, the best of a bad set of models is still a bad model. Some measure of the absolute performance of the models should therefore be included in any analysis considering multiple models (or single models for that matter).

A recent paper criticised the tendency for ecology journal papers to apply model selection techniques without assessing the absolute performance of the ‘best’ model [12]. They found that, out of 119 papers considered, only 55 included some measure of the absolute goodness of fit (the R^2 , the adjusted R^2 or chi-squared tests, for example) of the best model. They suggest that, as a matter of course, when statistical modelling approaches such as generalised linear models are used, a ‘null’ model should be included in the model selection process to help determine whether the ‘best’ model really does provide a better fit to the data. In fact, in this deliverable, it is argued that this benchmarking approach should be taken further and that benchmark models needn’t consist of simply a null model but that a general approach can be taken. This also means that benchmark models should be considered even when no obvious ‘null’ models exists, e.g. when the model of interest is physics based rather than statistical.

In this chapter, benchmark models, formed using historical data only, are defined and it is suggested that these should be treated as additional candidate models. Since, with enough data, it is always possible to build robust benchmark models, these provide a minimum level of skill required for a set of forecasts to be considered informative. If none of the model based forecasts considered can be found to outperform these benchmark models on average, they are not considered to be useful since it would be more informative to issue a forecast based on the benchmark model instead.

Although it is certainly the case that some measure of the absolute performance of the ‘best’ model should be calculated, in practice, more care is needed than, say, testing whether that model performs significantly better than the best benchmark model. This is because, whilst it is true that one can simply apply a single significance test of the absolute fit of the ‘best’ model, this does not remove the problem of multiple testing since that model has already been identified as one that performs relatively well in some sense (i.e. through model selection). This means that the probability of a type I error, i.e. wrongly rejecting the null hypothesis that the ‘best’ model does not provide a useful fit to the data, will be increased above the selected significance level. This greatly increases the chance of selecting a non-informative or even misleading model for future decision making. This problem is discussed in more detail in chapter 5 in which a ‘sanity-check’ test is defined to estimate the probability that the model selection statistic of the ‘best’ model occurred by chance.

This document is mostly concerned with probabilistic forecasts and thus, for continuous variables (although animal populations are not continuous, it is often convenient to treat them as such for modelling purposes), each forecast consists of a probability density function (PDF). In the ecological modelling examples described at the beginning of this document, for example, a generalised linear model approach is used to predict the relative population change (defined in the appendix). As such, whilst forecasts from such models are often interpreted as point forecasts with some uncertainty attached to them, they are better interpreted as probabilistic forecast distributions. In fact, it is already common to interpret the output of generalised linear models as probabilistic forecasts since information criteria, which are extremely common as measures of relative model performance, require the likelihood as an input and thus the underlying forecasts to be probabilistic in nature.

Two types of benchmark model are described in the following section: climatology, which is simply a distribution of past states, and persistence forecasts, which use the previous observation as a basis for a forecast of the next. Given enough past data, it is always possible to produce forecasts from benchmark models. As suggested in [12], it is usually sensible to include a ‘null’ model when appropriate. Often, in fact, the null model can be interpreted as either a specific form of the climatology (e.g. past states fitted with a Gaussian distribution) or as a specific form of persistence forecasts. The performance of a set of forecasts from each candidate model can be stated relative to that of the best benchmark model. This provides a measure of the absolute performance of the models since, if the models cannot provide more information than the benchmark models, they are of little use in practice. When the forecasts are evaluated probabilistically using the mean ignorance score (described in appendix B.3), the score of the forecasts relative to that of the best performing benchmark model can be interpreted as the mean number of bits of information gained from using the former rather than the latter.

4.1. Climatology

A climatology is a distribution of past observed states over some relevant time period that provides a useful summary of the probability of different outcomes given no other relevant information. Consider, for example, a case in which a temperature forecast is required at Heathrow airport five days into the future. Modern weather forecasting techniques can generally be expected to give informative probabilistic predictions over this time scale. At a lead time of say 60 days, however, weather models generally can *not* be expected to give informative probabilistic predictions. An alternative approach is to consider a climatology, a distribution of temperatures at Heathrow airport on or close to that calendar day over previous years. A climatology can thus be considered as an alternative, albeit fixed, probabilistic forecast. The longest lead time at which weather forecasts can be expected to outperform the climatology gives an indication of how long into the future useful forecasts can be made. Note that, in this example, the climatology was conditioned on the day of the year of interest. In many cases, however, there is no obvious seasonality in the data and thus all past observations can be considered. This is the case for annual counts of animal populations, for example.

Since a climatology is based on a set of past observations, it is necessary to adopt some approach with which to convert those observations into a probability density function. One approach would be to assume that the observations follow some parametric distribution such as a Gaussian distribution and to estimate the parameters from the data. In many cases, however, there is no underlying reason to choose any particular parametric distribution. In this case, an alternative, non-parametric approach can be applied. One such approach is kernel density estimation [13]. In kernel density estimation, each data point is replaced with a 'kernel', which is a probability density function centred on that data point with some parameter that governs its dispersion. The estimated density function at any given point is then the average of the density of those kernels. It is common to use a Gaussian kernel centred around each observation with the standard deviation of the kernel governing its width. The remaining choice to be made is that of the value of the standard deviation of the kernel, or the 'kernel width'. One approach to the selection of the kernel width is to use leave-one-out cross validation, maximising the average performance of the distribution as a forecast of the remaining point. The performance can be measured using the ignorance score, for example.

4.2. Persistence

When observed outcomes form a correlated time series, an alternative benchmarking approach is to use persistence forecasts, that is forecasts based purely on the previous observation. Whilst, in the context of point forecasts, a persistence forecast of the next observation is simply the last observation taken, in probabilistic forecasting, some approach is required to form a probability density function based around it. A simple approach is to create Gaussian forecast distributions with constant variance centred around the previous observation. Under this approach, forecasts take the form

$$f(n_i) = N(n_{i-1}, \sigma^2) \quad (4.1)$$

where n_i is the i th scalar observation of a time series and σ is the standard deviation of the forecasts. The value of σ is chosen to optimise the average performance of the forecasts over some training set given some method of evaluating forecast performance such as the ignorance score.

4.3. Null Models

Although the form of climatological and persistence forecasts described above can provide useful benchmark forecasts, sometimes, forecasts of these kinds arise naturally from 'null' models which are statistical models configured to have no input variables. For example, generalised linear models such as the stochastic Ricker and Gompertz models can be applied with no predictor variables, i.e. with only a constant term. Such a model can provide an alternative approach to calculating either a climatology or a set of persistence forecasts. For example, for the Stochastic Ricker and Gompertz models, if the predictand of interest is the relative population change (as opposed to the actual population), the null model is equivalent to fitting a climatology with a Gaussian distribution. In experiments involving statistical models such as these, as well as considering climatology and persistence (when appropriate), it is also useful to consider null models as potential benchmark models.

4.4. Example- Benchmarking Ibx Population Forecasts

In the ibex population modelling example described in appendix B.5, the aim is to predict the relative population change from one year to the next. In the original paper from which the models are taken, modified Stochastic Gompertz and Ricker models were fitted with various combinations of the three candidate variables and Akaike's information criterion was calculated for each. In this section, the results of forming benchmark models for this example are presented.

In this example, the predictand is taken to be the relative population change $R_i = \log\left(\frac{n_{i+1}}{n_i}\right)$ and the model selection statistics are calculated based on forecasts of this quantity (rather than the populations themselves). A climatology can thus be constructed using past observations of this variable. A ‘null’ model distribution can also be formed using the Stochastic Ricker or Gompertz Model with all parameters except the intercept and the variance set to zero. The null model therefore takes the form $R_i \sim N(a, \sigma^2)$ where a and σ are parameters to be selected. As previously mentioned, this approach simply fits a Gaussian distribution to past observations and is thus equivalent to fitting the climatology using a Gaussian distribution. A climatology is also calculated using the kernel density estimation approach described in section 4.1. The estimated climatologies calculated using kernel density estimation (blue) and by fitting a Gaussian distribution (red), that is using the null model, are shown along with the positions of the past observations from which they are formed in figure 4.1. Here, the difference between the two approaches becomes clear. Kernel density estimation has much greater flexibility in terms of the shape of the distribution than the null model that constrains the distribution to be Gaussian.

Annual population counts are, by their nature, usually highly correlated since it is often the same animals that are counted from one year to the next. As such, if the predictand of interest is the actual population counts, it is worthwhile to produce persistence forecasts. In this example, however, the predictand is the relative population change R_i . It is unclear that the population *change* should be autocorrelated in a way in which a persistence forecasting approach would be effective. This, of course, can easily be tested by calculating the autocorrelation of the time series of relative population changes. After doing this, it is found that there is an autocorrelation of -0.38 with a p-value of 0.018 , i.e. significant evidence of negative autocorrelation. It is thus clear that a persistence approach would not be effective and so forecasts of this kind are not considered in this case. Persistence forecasts are, however, considered in chapter 6, which is concerned with modelling the wild reindeer population in Hardangervidda, since the predictand of interest here is the actual population size.

In order to determine which of the two candidate benchmark models performs best, model selection statistics can be calculated for each. The one with the best statistic can then be chosen as the model with which to compare the performance of the model based forecasts. The AIC and the cross-validated mean ignorance (with 90 percent resampling intervals) of each benchmarking approach are shown in table 4.1. Since both the AIC and the mean ignorance are far lower, here, the climatology formed using kernel density estimation likely provides a better benchmark model than the null model.

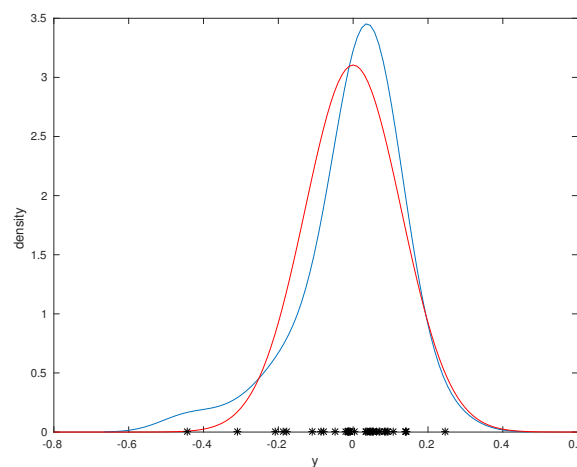


Figure 4.1: Estimated climatology for the relative population change of the ibex population in Gran Paradiso National Park calculated using kernel density estimation (blue) and a Gaussian distribution (red), that is using the null model. The black stars show the positions of the past observations of the relative population change.

Model	AIC	Ignorance
Climatology	-57.7	-0.93(-1.23, -0.41)
Ricker/Gompertz Null Model	-46.4	-0.77(-1.15, +0.26)

Table 4.1: AIC and cross-validated mean ignorance (with 90 percent resampling intervals) score of the climatology of the relative population changes formed using kernel density estimation and fitted using a Gaussian distribution (the null model).

In the paper from which this example is taken, each of the models were compared using the AIC. Having identified the best benchmark model as the climatology, it is now possible to express the performance of each of those models relative to this. This is done by simply subtracting the AIC of each candidate model from that of the climatology. If the resulting number is negative, then that model has more support than the benchmark model. Although, in the original paper, cross-validation

was not considered, here the models are also compared using leave-one-out-cross-validation with the mean ignorance score as the evaluation method. Whilst these two approaches are asymptotically equivalent, they will usually be expected to give a different ordering of models for finite sample sizes. Similarly to the AIC, the cross-validated mean ignorance score is expressed relative to that of the climatology in each case. The mean ignorance score has an advantage over the AIC in that, when used to compare two forecasting systems (in this case, a candidate model and the climatology), it has a clear interpretation in terms of how much probability density is placed on the outcome, on average (see appendix B.3 for details).

The results of the model selection process for this example are shown in table 4.2. Columns headed by *b*, *c* and *e* indicate whether density dependence, snow cover and the interaction between the two have been included in the model respectively. It is found that almost all of the models outperform the climatology under both measures. The difference in mean ignorance can be interpreted as the mean number of bits of information gained from using the model rather than the climatology. For example, the best model, model 11, has a mean ignorance relative to climatology of -0.82 , implying a gain of 0.82 bits, that is model 11 places $2^{0.82} = 1.76$ times more density on the outcome than the climatology, on average. For the cross-validated mean ignorance, 2.5 – 97.5 percent resampling intervals of the mean are also given. Since the intervals do not contain zero, four of the models perform significantly better than the climatology at the 5 percent level. Care, of course, needs to be taken when evaluating the significance of models. The question of how to assess whether the results of the model selection process occurred by chance is addressed in chapter 5.

Model	b	c	e	Parameters	DD Type	AIC	Mean Ignorance
M1	*	*	*	4	n	-20.4	-0.34(-0.77, +0.57)
M2	*	*	*	4	x	-23.3	-0.46(-0.88, +0.28)
M3		*	*	3	n	-22.4	-0.41(-0.83, +0.50)
M4		*	*	3	x	-25.4	-0.51(-0.93, +0.25)
M5	*		*	3	n	-19.5	-0.40(-0.75, +0.14)
M6	*		*	3	x	-19.2	-0.39(-0.73, +0.02)
M7	*	*		3	n	-15.7	-0.32(-0.66, +0.08)
M8	*	*		3	x	-18.6	-0.38(-0.71, +0.02)
M9			*	2	n	-10.8	-0.28(-0.62, +0.03)
M10			*	2	x	+0.96	-0.01(-0.26, +0.39)
M11	*	*	*	7	n	-39.2	-0.82(-1.54, -0.27)
M12	*	*	*	7	x	-38.5	-0.79(-1.46, -0.28)
M13		*	*	5	n	-34.4	-0.69(-1.25, +0.02)
M14		*	*	5	x	-34.2	-0.67(-1.18, +0.01)
M15	*		*	5	n	-25.5	-0.50(-0.99, -0.10)
M16	*		*	5	x	-23.1	-0.42(-0.82, -0.04)
M17	*	*		5	n	-13.1	-0.11(-0.52, +0.64)
M18	*	*		5	x	-20.7	-0.36(-0.72, +0.08)
M19			*	3	n	-9.4	-0.21(-0.57, +0.24)
M20			*	3	x	+1.4	+0.04(-0.27, +0.82)

Table 4.2: AIC and cross-validated mean ignorance scores (with 2.5-97.5 percent resampling bars) expressed relative to the climatology of each ibex model. Columns headed by *b*, *c* and *e* indicate whether density dependence, snow cover and the interaction between the two have been included in the model respectively. Also shown are the total number of free parameters and the form of density dependence used. The model with the best AIC and mean ignorance score is highlighted in blue.

5. A Sanity Check For Model Selection

As discussed at length in chapter 4, standard model selection techniques only compare the relative performance of a set of models and thus, if none of those models are effective, the model selection procedure becomes redundant. A benchmarking approach, in which simple statistical models based on past observations are used as candidate models, can help determine the absolute performance of a set of forecasts. This alone, however, is not enough to distinguish genuine forecast skill from random chance since, the more models that are tested, the higher the probability of finding a model that outperforms the best benchmark model by chance.

The methodology in this chapter aims to answer a simple question: given the models tested, what is the probability that the model selection statistic of the ‘best’ model occurred by random chance? Using a simple permutation test approach, the sanity-check does not take into account any of the benchmark models and therefore cannot be interpreted as an indicator of the significance of the forecasts relative to the benchmark model. It is rather a simple procedure to help reassure the modeller that, as long as the estimated probability is low, their results are unlikely to have occurred by chance, i.e. through testing a large number of different models. It is argued that, when computationally feasible, this information should accompany every model selection procedure to reassure the modeller that the inputs to the best model really do contain useful information about the subject of interest.

5.1. Permutation Tests

A permutation test is a nonparametric statistical test in which the significance of a test statistic is obtained by calculating its distribution under all different permutations of the set of observed outcomes. For example, a permutation test for the slope parameter of a simple linear regression would be performed by permuting the positions of the y values (the dependent variable), keeping the x values (the independent variables) in their original positions and calculating the slope parameter under all possible combinations of y . The position of the slope parameter that has been calculated from the data in their original positions would then be compared to this distribution to calculate a p-value. In practice, it is often computationally prohibitive to consider all possible permutations and thus permutations are randomly chosen a fixed number of times. Such tests are called randomised permutation tests. Permutation tests have a number of advantages over standard parametric tests. Unlike the latter, no assumptions about the distribution of the test statistic under the null hypothesis are required since the method draws from the exact distribution. Permutation tests thus give an exact test and, as such, randomised permutation tests are asymptotically exact.

5.2. Sanity-Check Test

Instead of testing the significance of individual models, a test for the entire model selection procedure is proposed. Using a randomised permutation test, the probability that the model selection statistic of the best performing model could have occurred by chance is calculated. This is done by randomly permuting the observed outcome values, calculating model selection statistics, and recording the statistic of the best performing model. This process is then repeated a large number of times to gather draws from the distribution of best performing model statistics under the null hypothesis that none of the variables are effective predictors of the outcomes. The position of the observed model selection statistic (i.e. that calculated from the data in their original order), with respect to this distribution is then calculated to estimate the probability of observing a model selection statistic as good, or better by chance.

Define a set of outcomes $\mathbf{y} = y_1, \dots, y_n$ and a matrix of potential predictor variables $X = \mathbf{x}_1, \dots, \mathbf{x}_n$ where \mathbf{x}_i is a vector of predictor variables corresponding to the i th outcome y_i , that is $\mathbf{x}_i = x_{i,1}, \dots, x_{i,d}$. Suppose that each model includes as an input some subset of the possible combination of the variables in \mathbf{x} . Given the candidate models, the distribution of the model selection statistic of the ‘best’ model is sought under the null hypothesis that each of the input variables is independent of the outcome. Since it is true both that each variable can be present in different candidate models and that different candidate variables can be correlated, the candidate models cannot be considered to be independent of each other. It is therefore important that these dependencies are preserved in attempting to find the distribution of the null hypothesis. A solution to this problem that does all of these things is to permute the outcomes \mathbf{y} whilst ensuring that none of the outcomes fall into their original positions. Therefore, the test is applied using the following procedure:

1. Calculate and record the model selection statistics of the ‘best’ model M_{obs} using the outcomes in their original order.
2. Set $j = 1$
3. Randomly permute the outcomes ensuring that none of them fall into their original positions.

4. Calculate the ‘best’ model selection statistic over all models M_j with the permuted data.
5. Set $j = j + 1$
6. Repeat steps two to four until $j = J$.
7. Calculate $p_{obs} = \frac{1}{J} \sum_{j=1}^J I(M_j < M_{obs})$ where $I(M_j < M_{obs})$ is the indicator function.

The test described above is demonstrated on both the ibex population modelling example and the reindeer population study described in chapter 6.

5.3. Ibex Population Models in Gran Paradiso National Park

The sanity-check test that was introduced in section 5.2 is now demonstrated on the ibex population modelling example described in section B.5. In the experiment outlined in the original paper, simple modified stochastic Ricker or Gompertz models with different variations of variables were fitted to the population data. A total of 20 combinations of variables were assessed and compared using Akaike’s Information Criterion (AIC). The sanity-check test is now applied to these models to estimate the probability that the model selection statistic of the ‘best’ model occurred by chance.

Since each of the models is computationally undemanding, it is possible to run the sanity-check test with a large value of J , the number of permutations of the original data set. Applying the test, no smaller AIC values were found than the best model from the observed data. The positions of the smallest AIC over all 20 models from each permutation along with their CDF are shown in the top panel of figure 5.1 along with the AIC from the original ordering of the data (all relative to the AIC of the climatology). The equivalent, but with the mean ignorance rather than the AIC, is shown in the lower panel. From, these results, it is clear that it is extremely unlikely that the ‘best model’ in the the model selection procedure occurred by chance and thus confidence can be held in the results.

Interestingly, out of the resampled data sets, the ‘best’ model outperforms the climatology (the best benchmark model) in only 1.00 percent of resampled cases for the AIC and 1.81 percent of cases for the cross-validated mean ignorance (N=65536).

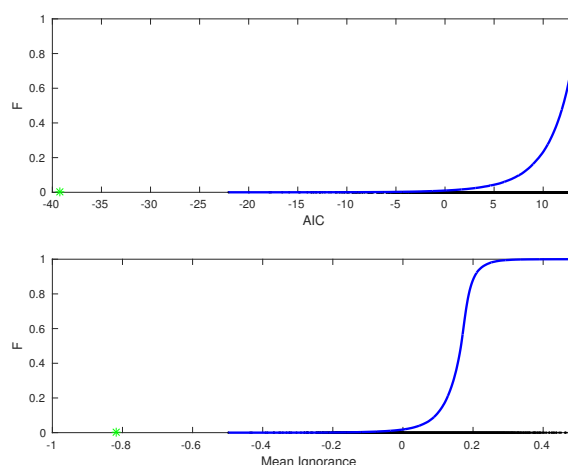


Figure 5.1: Top: Smallest AIC values from resampled data (black dots), their CDF and the smallest AIC from the observed data (green star) for the ibex example. Bottom: the same for cross-validated mean ignorance

6. Modelling the Wild Reindeer Population of Hardangervidda

This chapter presents some of the research featured in an upcoming paper (Bargmann et. al, 2018) studying the population dynamics of the wild reindeer population of Hardangervidda national park, which will be submitted in early summer 2018 as part of the Ecopotential project. The authors of that paper are Tessa Bargmann (University of Bergen), Ed Wheatcroft (London School of Economics), Simona Imperio (Consiglio Nazionale delle Ricerche (CNR), Italy) and Ole Reidar Vetaas) and credit goes to them for them for the work in this chapter. All of the words in this chapter, however, have been written by Ed Wheatcroft.

Whilst the focus of the aforementioned paper was mostly on the ecology of the population, the aim of this chapter is to focus more on the methodology that was used. Much of the discussion on ecology is therefore left out but these details can be found in the paper. This chapter thus focuses on why certain methodology was used and provides deeper insight into the statistical results.

6.1. Introduction

Hardangervidda National Park in central Norway supports a large wild reindeer population. The aim of this work is to attempt to understand the factors that affect the size of this population. This is done using a population modelling approach in which various combinations of factors are considered as predictors of the population at the next count. Whilst detailed discussion of why they are considered are left for the upcoming paper, the following climatic factors are considered as potential predictors of the population:

- The current size of the population (density dependence).
- Mean temperature over January and February.
- Mean temperature over July and August.
- The number of summer growing degree days from June to September (days above 5 degrees celsius)
- The proportion of the population hunted and killed.

6.2. Methods

6.2.1. Data Sources and Preparation

Calf counts and data on the population structure of the wild reindeer population from 1994 to 2015 were provided by the Wild Reindeer Centre. Since population structure counts only cover a subset of the population, these were combined with the calf count data to estimate the total population size. Gridded datasets of the mean daily temperature and the total accumulated precipitation at 1x1km resolution were provided by the Norwegian Meteorological Institute (version 1.1). This dataset was produced by interpolating data from nearby weather stations and using a Digital Elevation Model at a 100m resolution. The Norwegian Water Resources and Energy Directorate (version 2.0.1) provided daily snow data. Snow data is calculated using snow models run using weather data from the Norwegian Meteorological Institute (for more information see: Saloranta, 2014). All climate data was clipped to include only the known extent of the Hardangervidda reindeer population and then summarized for each year and season, depending on the variable in question.

6.2.2. Population Models and Forecasting

The aim of this research is to investigate the effects of density dependence, hunting and various climatic conditions on the wild reindeer population of Hardangervidda national park. Firstly, the effect of each potential predictor variable on the relative population change is assessed individually. A formal statistical test is then applied to assess whether there is a significant linear relationship between the variable in question and the relative population change. Secondly, a probabilistic forecasting framework is defined in which various combinations of variables are tested as predictors of the population at the next count. Whilst the former approach allows for the testing of single predictor variables, the latter allows for combinations of these variables and interactions between them.

In this research, *probabilistic* forecasts are constructed. There are very good reasons to take this approach. As described in section 6.2.1, the population counts for each year can only be estimated from population structure data. This means that, in practice, some uncertainty as to the true population size will always exist in practice. This means that, even if a perfect deterministic model of the population dynamics were to be in hand, perfect point forecasts of the following year's

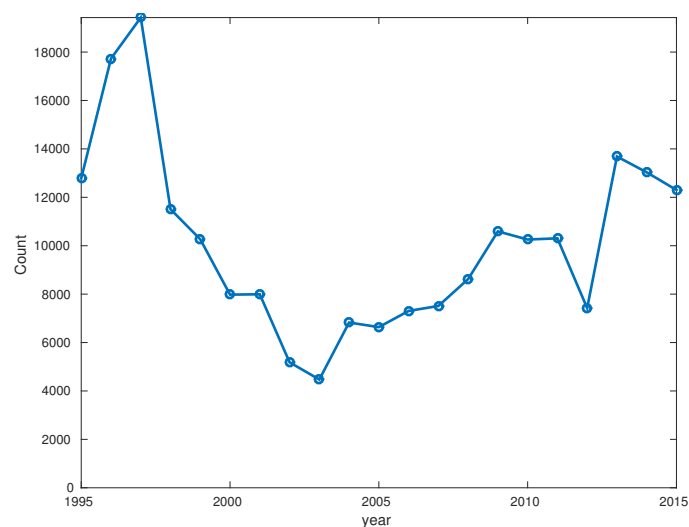


Figure 6.1: Estimated reindeer counts from 1994 to 2015 in Hardangervidda.

population would not be attainable due to the observational uncertainty in the current year's population. In addition, many population models have a stochastic element and thus probabilistic forecasts arise rather naturally.

6.2.3. Testing For Density Dependence and Other Drivers of Population Change

A number of different variables are considered as potential drivers of the reindeer population. Each of these relationships is assessed firstly by simply plotting the value of each variable against the relative population change and, secondly, by applying a formal statistical test. The statistical test applied, Pollard's randomisation test [14], was first designed to test for density dependence, i.e. to test whether the size of the current population linearly affects the relative population change. It is easily extended to test the effects of other variables as well, however. It is worth noting that the test only considers linear relationships and, by design, will not detect some nonlinear relationships. Whilst it may be possible to apply some test that detects some nonlinear relationships, in practice, the small sample size would mean that, if any such relationship does exist, it would likely not be picked up by the test (except, perhaps, something like a simple quadratic relationship). It is thus useful to plot the data so that any obvious nonlinear relationships between the variables can be identified. In fact, due to the small sample size (21 years of population data), the power of any statistical test will be low. It is thus advisable not to over interpret the results. A lack of a significant relationship should not be interpreted as strong evidence that such a relationship does *not* exist, for example.

6.2.4. Density Dependence

Large population sizes can often cause increased mortality in animal populations due to the increased competition for scarce resources. It is generally held that, for any habitat, there is a limit on the size of the population. To see whether there is obvious evidence of density dependence in the reindeer population, the current population size is first plotted against the relative population change. A formal test is then applied to attempt to detect density dependence. A large body of literature has been devoted to deriving tests for density dependence. Many of these make assumptions that can be problematic in practice, however. For example, Bulmer's auto regression test [15] assumes the existence of a trend in the data which, in practice, can be difficult to justify. A linear regression approach, in which the slope parameter is tested for significance, can also be problematic in practice because the standard t-test assumes that residuals are iid Gaussian. This can be hard to prove in practice, particularly with a small data set. In addition, this approach can be misleading when the data have a trend [16] (recall that, for Bulmer's Test, the opposite is true). An alternative approach, which makes fewer assumptions, is Pollard's randomisation test [14]. For this test, each observed relative population change (the response variable) is randomly assigned to an observed population so that any underlying association is (almost certainly) removed. The regression slope parameter is then estimated using least squares estimation. This process is then repeated many times (one hundred thousand in this case) and the estimated parameter of the slope is recorded. The position of the estimated slope parameter from the actual data is then calculated relative to the ordered set of values from the randomised data. This allows a p-value to be found. For example, if the estimated slope parameter from the non-randomised data is larger than exactly 95 percent of the values from randomising, it is given a p-value of 0.05 (or 0.1 if a two sided test is used). The advantage of this approach is that it is distribution free, that is, unlike the t-test, no assumption is made about the distribution of the residuals. In addition, any inflation (or deflation) of the slope parameter resulting from an underlying trend will also be present in the randomised data sets and thus the risk of wrongly rejecting the null hypothesis (that the

data are uncorrelated) is removed. It can be shown, however, that tests of this kind can have a true type I error rate far from the stated α but that, by studentising the variables (i.e. dividing by their standard deviation), the test is asymptotically exact [17]. This approach is taken throughout the paper. Of course, this test, and most other well known tests, only test for linear density dependence. Were some nonlinear relationship to exist, standard tests would not necessarily detect this. This is why it is important to plot the data to rule out any obvious underlying nonlinear dependencies that are not easily described parametrically.

In addition to testing for density dependence, it useful to test how other variables relate to the relative population change. The same randomisation test used to test for density dependence can be used with the current population replaced with some other variable. Scatter plots with their least squares linear line of best fit are also shown to help determine whether any underlying nonlinear relationship is evident. Cook's distance [18] is calculated in each case to test for influential observations. An observation is considered 'influential' if its Cook's Distance exceeds one.

6.2.5. Benchmark Models

6.2.5.1. Climatology

As described in appendix 4.1, the climatology is defined as a distribution derived purely from past observations, in this case population estimates. If no useful information (including the last population size) exists regarding the next estimated population size, the climatology can be considered the best possible probabilistic forecast, as long as it is a robust estimate of the underlying distribution of possible population sizes. As a result, if a model based probabilistic forecast does not outperform the climatology, on average, it is of little value since a better score could be achieved by simply by issuing the climatology as the forecast instead.

The estimated climatology formed using the approach described in section 4.1 is shown in figure 6.2. Although the estimated climatology is bimodal, it is hard to say whether this pattern is 'real' or results from the limited sample size.

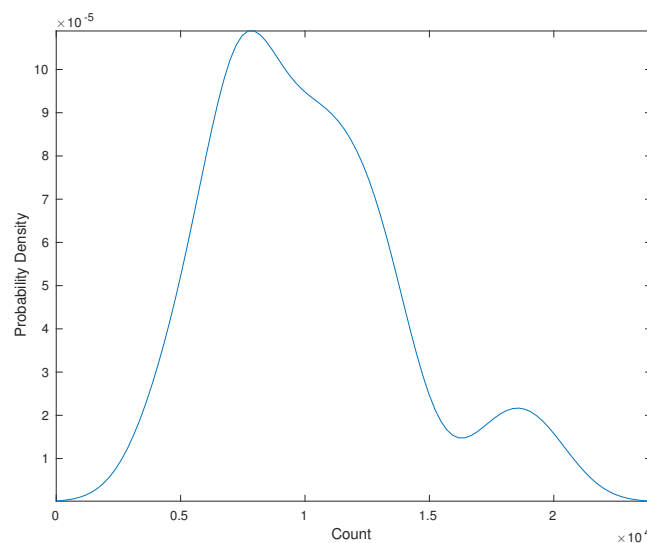


Figure 6.2: Estimated climatology of reindeer counts in Hardangervidda.

6.2.5.2. Persistence Forecasts

An alternative benchmark model can be formed on the basis that the population of reindeer will be largely similar to that of the previous year. The approach used to build persistence forecasts for this example is described in section 4.2. Since a population cannot be negative, the forecasts are renormalised such that the integral of each persistence forecast over $(0, \infty)$ is equal to one.

6.2.5.3. Ricker Null Model Persistence Forecasts

An alternative approach to forming persistence forecasts can be taken using the null Stochastic Ricker Model in which only the constant term is used, i.e in the form

$$n_{i+1} = n_i \exp(a) + \sigma \epsilon, \quad (6.1)$$

where a and σ are parameters to be selected and ϵ is a draw from a Gaussian distribution. To form probabilistic forecast distributions, ensembles consisting of 10000 members each are defined by taking different randomly drawn values of ϵ for each member. Kernel density estimation is then used to form forecast distributions with the bandwidth chosen to minimise the mean ignorance score using leave-one-out cross validation.

6.2.6. Model Selection

It is certainly the case that the reindeer population estimates are correlated and thus care needs to be taken in terms of model selection. However, the inputs to the models can generally be assumed not to be correlated and thus the extent to which the forecasts are dependent is reduced. In fact, the only case in which the inputs to the models can be considered non-independent are when density dependence is considered. Therefore, both cross-validation and AIC_c are used in their original form but care should be taken in the interpretation of the results, particularly from models that include density dependence.

6.3. Results

6.3.1. Choosing a Benchmark Model

The AIC and the mean cross-validated ignorance scores (with 95 percent resampling intervals) for each benchmark model of the reindeer population are shown in table 6.2. Out of the benchmark models, the Stochastic Ricker Persistence performs the best under both the AIC and the mean cross-validated ignorance and is thus used as the zero value in both cases, i.e. the AIC and mean ignorance are subtracted from each of the other models.

Model	AIC	Ignorance
Climatology	382.1	14.03 (13.67,14.69)
Persistence	377.0	13.72 (13.17,15.17)
Ricker Null Model	370.7	13.43 (12.93,14.39)

Table 6.1: AIC and cross-validated mean ignorance (with 95 percent resampling intervals) scores for benchmark models of the reindeer population in Hardangervidda.

6.3.2. Testing For Density Dependence

The population for each given year is plotted against the relative population change (where known) over the following year in figure 6.3 along with the least squares linear line of best fit. Although only linear density dependence is formally tested, there is no obvious sign of a nonlinear relationship either. Applying a two-sided randomisation test, a p-value of 0.113 is found, giving a low level of significance. Although there is only low evidence of linear density dependence, more investigation regarding the existence and nature of density dependence would be useful, particularly given the emergence of new data points in the future.

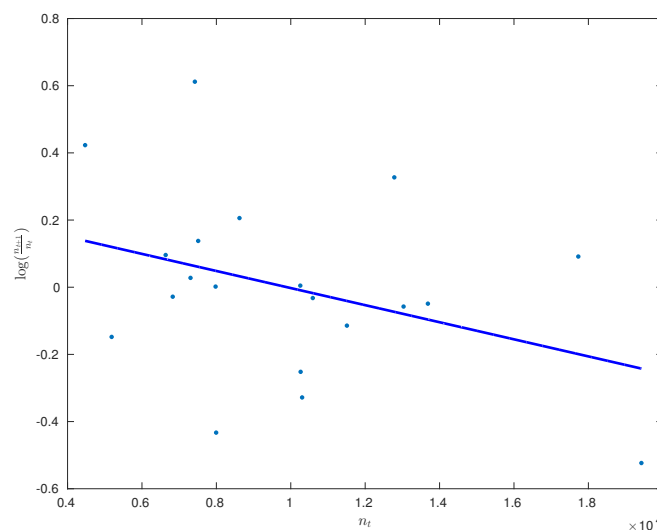


Figure 6.3: Scatter plot of population size against relative population change along with the least squares linear line of best fit. Here, there is low evidence here to suggest density dependence in these data.

6.3.3. Effects of Hunting

In Hardangervidda, each year, a number of hunting licenses are issued so that, for a given price, the holder is entitled to shoot and kill one reindeer within the limits of the national park. Each time an animal is successfully killed, this is reported back to the park authority. It is believed that this reporting yields a reliable estimate of the number of animals killed through hunting. Hunting takes place during the month of August and September which is after the annual population counts take place. Intuitively, it is to be expected that the number of animals hunted in the autumn of some year i would be expected to have the greatest impact on the population change between year i and year $i + 1$. The proportion of the total population reported killed in year i is plotted against the relative population change in figure 6.4 along with the least squares linear line of best fit. Here, the slope of the line is negative which one would expect if hunting activity tends to reduce the following year's population. The randomisation test yields a p-value of 0.109. In this case, however, it makes sense to apply a one-sided test since hunting is not expected to have a positive effect on the population. The p-value in this case would thus be 0.054, yielding some evidence that hunting affects the population negatively. To gain stronger insights into the true effect of hunting on the population, more data would be required.

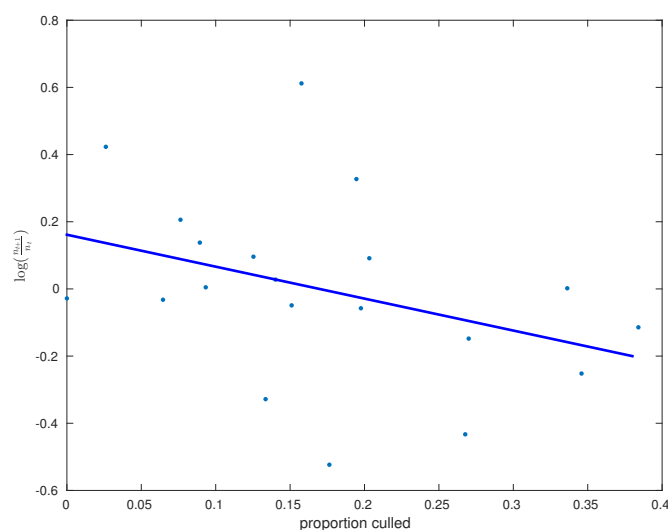


Figure 6.4: A scatter plot of the proportion of the population reported killed against the relative population change with its least squares linear regression line. As would be expected were hunting to have a negative effect on the population, the slope of the regression line suggests that higher hunting activity tends to reduce population growth. A relatively high p-value suggests only weak evidence for this, however.

6.3.4. Climate Effects

A number of different climatic variables are considered as potential predictors of the relative population change and therefore the population itself. These data are generally derived from gridded reanalysis data averaged over the known range of the reindeer.

6.3.4.1. Winter Temperature

One potential driver of changes in the reindeer population is snow cover affecting the ability of the animals to be able to reach lichen on the ground. Snow cover itself is not necessarily a problem since reindeer are able to dig through up to 150cm of snow. When the snow melts, however, the snow turns to water and then refreezes. This creates a layer of ice which can cause problems for the animals by making it hard to break through to the lichen on the ground.

Given the data that are available, it is difficult to quantify the mechanism in which the snow melts and refreezes. When the mean temperature stays far below zero, however, it seems less likely that the snow will melt and refreeze. The icing effect will therefore be less likely to occur. In cold temperatures, precipitation is also less likely to fall as rain that can freeze and cause the same icing effect as melting snow. The mean temperature over the winter months is therefore a possible driver of population change. The mean temperature over both January and February is plotted against the relative population change in figure 6.5 as well as the least squares linear line of best fit. Here, there appears to be a negative linear relationship between the mean temperature and the relative population change. This suggests that colder temperatures tend to result in lower mortality in the population. The point denoted with a green star is an influential observation (according to Cook's distance) which corresponds to population changes over the year from 2009 to 2010. It is not clear whether this point occurs because the linear relationship does not extend to such low temperatures or whether it is caused by

some confounding factor in that year (for example, in 2016, over 300 reindeer were struck and killed by a single bolt of lightning). Without an obvious explanation, it is difficult to justify extrapolating the regression line to include this point. Applying Pollard's randomisation test, a p-value of 0.0142 is found and, when the year 2009-2010 is left out of the analysis, the p-value is 0.0005, indicating very strong evidence that the winter temperature has a negative effect on the population in a typical winter.

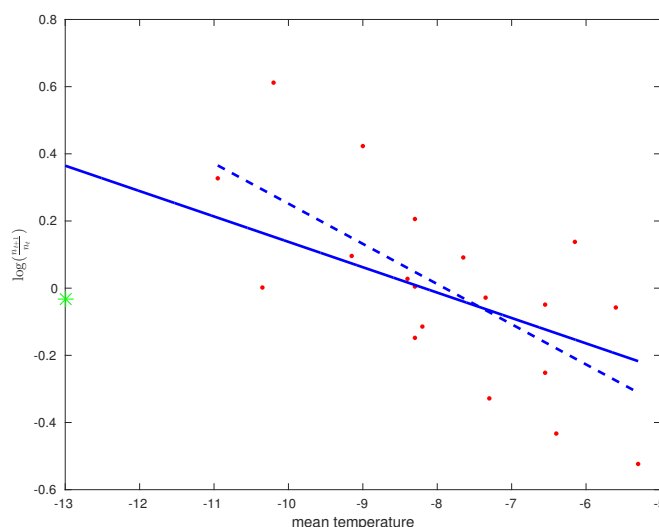


Figure 6.5: Plot of the mean temperature over January and February in year $i + 1$ against the relative population change $\log(\frac{n_{i+1}}{n_i})$. The green star corresponds to the year 2009-2010 which is an influential observation. The solid and dashed lines represents the least squares linear lines of best fit calculated with and without 2009-2010 respectively.

6.3.4.2. Winter Snow Days

Since reindeer can only dig through a limited amount of accumulated snow, one might expect the number of winter snow days (days in which it snows from January to March) to affect the population. The number of winter snow days is plotted against the relative population change in figure 6.6 along with the least squares linear line of best fit. Here, there appears to be little evidence of a relationship. The p-value from the the randomisation test is not significant at any standard level ($p = 0.188$).

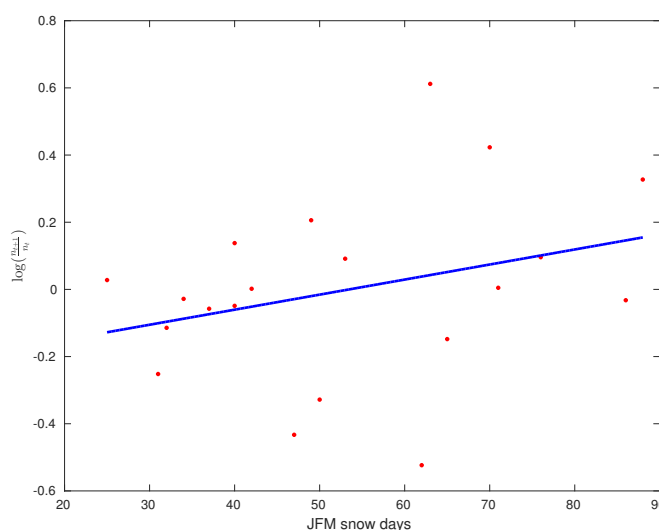


Figure 6.6: The number of winter snow days against the relative population change with the linear line of best fit. There is little evidence of a relationship here.

6.3.4.3. Summer Temperatures

The bodyweights of reindeer are believed to be negatively impacted by insect harassment in the summer. Lower bodyweights can then cause higher mortality in the following winter. Insect harassment tends to be greater during relatively

warm summers. The average summer temperature can thus be used as a potential predictor variable of the relative population change. The mean temperature over July and August temperature in the i th year is plotted against the relative population change between the years i and $i + 1$ in figure 6.7 along with the least squares linear regression line. Although the regression line has a negative slope, as would be the case if insect harassment does indeed negatively affect the population, the p -value for the randomisation test is not significant at any standard level ($p = 0.1685$).

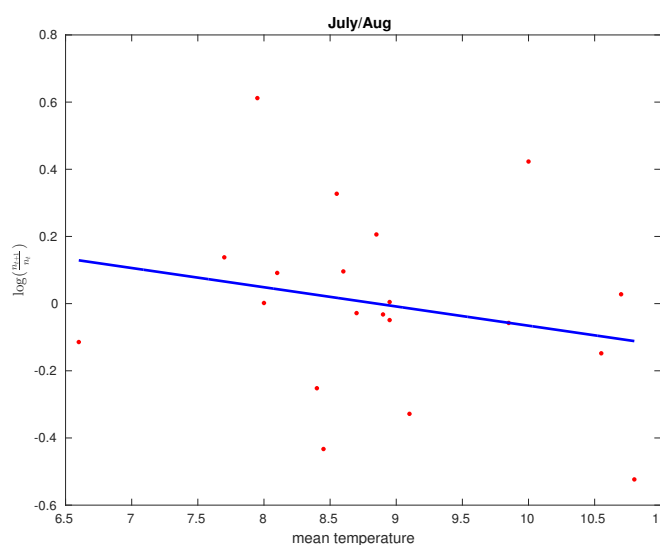


Figure 6.7: Plot of the mean temperature over July and August in year i against the relative population change $\log(\frac{n_{i+1}}{n_i})$. The line represents the least squares linear line of best fit.

The reindeer population is also thought to be affected by the quality and quantity of lichen in the summer months. Good quality lichen allows the animals to be suitably healthy going into the winter season and thus may be expected to decrease winter mortality. The number of growing degree days during the summer may therefore be expected to have some effect on the population. However, the number of growing degree days and the mean summer temperature are closely related and thus, if high summer temperatures genuinely have a negative effect on the population through insect harassment, this effect may be partially cancelled out if the number of growing degree days has a positive effect through improved quantity and quality of lichen. In figure 6.8, the number of growing degree days over the summer in year i is plotted against the relative population change from year i to $i + 1$ along with the least squares linear line of best fit. Here, there is weak evidence of a linear relationship ($p = 0.0733$). The sign of the slope parameter shows that the number of growing degree days and the relatively population change are negatively correlated (though there is only weak evidence of a robust underlying relationship). If indeed it is true that summer temperatures have conflicting positive and negative effects on the population, the overall relationship may be highly nonlinear and extremely difficult to capture given the small sample size in the study.

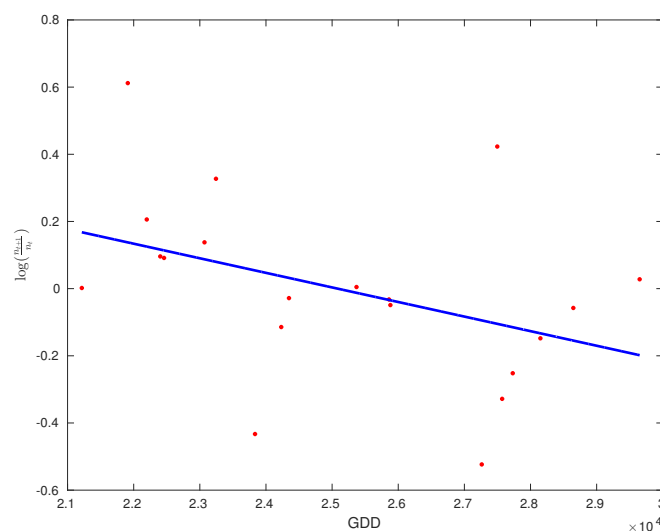


Figure 6.8: The number of growing degree days in year i against the relative population change with the linear line of best fit.

6.3.5. Model Selection

The AICc, the mean cross-validated ignorance scores (with 90 percent resampling intervals) and the R^2 for each candidate model are shown in table 6.2. As determined in section 6.1, out of the benchmark models, the stochastic Ricker null (persistence) model performs the best according to both the AICc and the mean cross-validated ignorance and is thus used as the zero value in both cases, i.e. the AICc and mean ignorance are subtracted from each of the other models. Any negative AICc or mean ignorance score thus outperforms the best null model on average.

The results in table 6.2 are contradictory both in terms of the ordering of the models and in terms of which outperform the benchmark models. The best three models according to the AICc actually have positive mean ignorance relative to the best benchmark model (the null model) indicating that, according to this measure, they are, in fact, counterproductive, on average. Since zero always falls within the resampling intervals, none of the models significantly outperform the null model according to the leave-one-out cross-validation approach. A rule of thumb for comparing AICc values, on the other hand [19] is that differences in AICc between 3 and 7 indicate considerably higher support for the model with the lower AICc. Leaving aside issues with multiple comparisons for the moment, there is therefore an apparent contradiction here as well. This merits further investigation. Figure 6.5 shows how the winter of 2010 was far colder than the rest of the years in the data set and does not follow the linear relationship that appears to be present in the rest of the points. Leave-one-out cross-validation penalises the model strongly for the existence of this point since, when the parameters are fitted over all of the other points, the outcome falls a long way into the tail of the forecast distribution. The existence of this point is also heavily penalised by the AICc, although to a lesser extent, since the parameters are always fitted over all of the points. It is clear from figure 6.5 that, in fact, there is a great deal of information in the winter temperature though this is not apparent in the model selection statistics. On the other hand, the only information about how very low temperatures affect the population comes from one particular point and so it is very difficult to say much about this from this information alone. It is therefore useful to assess the performance of each model when the year 2009-2010 is removed and thus not attempt to assess the effects of very low temperatures. The results of the analysis are shown in table 6.3. Here, the best model according to both the AICc and leave-one-out cross validation features the mean temperature over January and February and the proportion of animals hunted, although the best model according to the former also features growing degree days whilst, for the latter, the model includes the mean temperature over July and August. There is therefore far less of a contradiction between the two model selection techniques. An interesting question is then whether the sensitivity (i.e. the high impact of one outcome falling in the tail of the forecast distribution) of cross-validation with ignorance is a strength or weakness of the method. Addressing of this question, however, is left for future work.

So far, some attention has been given to whether the performance of each of the models outperforms the best benchmark model. It was found that, under leave-one-out cross-validation, none of the models significantly outperformed the best benchmark model whilst, using AICc, with a simple rule of thumb, some of the models could be considered to have considerably more support. As discussed at length in chapter 5, however, when multiple models are considered, there is a very high chance of observing results like these. In this chapter, a ‘sanity check’ randomisation test for model selection was described in which the probability of finding model selection statistics as good or better is calculated by permuting the data set. Applying the test with 4096 different randomised surrogate data sets, the probability of finding the best observed AICc value or lower by chance is estimated to be 11.7 percent and the probability of observing the lowest mean ignorance score or lower by chance is estimated to be 23.7 percent. A CDF of the best AICc value for each randomised data set along with the position of the value of the lowest AICc for the observed data is shown in the top panel of figure 6.9. Here, it is clear that the probability of observing an AICc as low or lower than for the observed data is high. Note how, for every randomised data set, the model with lowest AICc always outperforms the best benchmark model. The ‘sanity check’ test shows that the results of this model selection procedure need to be treated with extreme caution. The same test can be applied to the model selection procedure with the year 2009 to 2010 left out. In this case, the probability of observing an AICc smaller than for the original data set is estimated to be 0.017 percent whilst the corresponding probability for leave-one-out cross-validation is estimated to be 0.015 percent. Figure 6.10 is the same as figure 6.9 but with 2009-2010 left out. Clearly, it is very unlikely that the best model occurred by chance and thus much greater confidence can be had in the results.

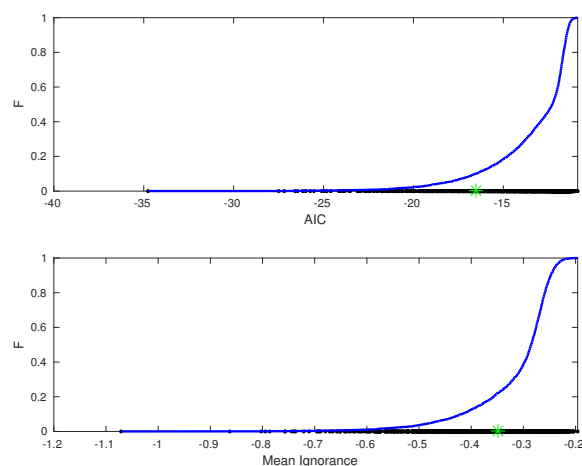


Figure 6.9: Top: Smallest AICc values from resampled data (black dots), their CDF and the smallest AICc from the observed data (green star) for the reindeer example. Bottom: the same for cross-validated mean ignorance

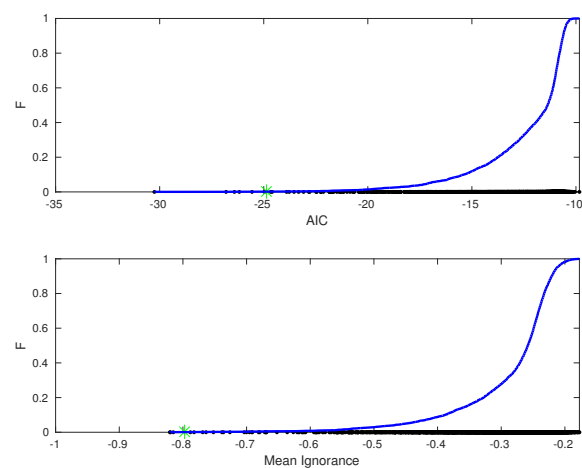


Figure 6.10: Top: Smallest AICc values from resampled data (black dots), their CDF and the smallest AICc from the observed data (green star) for the case in which the year 2009 to 2010 is left out. Bottom: the same for cross-validated mean ignorance

	JF	JA	GDD	DD	Hunted	DD Int	R ²	AICc	Mean Ign
M18	-0.057		-0.000		-0.850		0.454	-5.600	+0.076(-0.480, +2.058)
M15	-0.062	-0.069			-1.029		0.440	-5.334	+0.181(-0.565, +3.331)
M7	-0.070				-0.794		0.379	-5.273	+0.172(-0.526, +2.298)
M1	-0.076						0.285	-4.447	-0.015(-0.440, +1.051)
M16	-0.064	0.000					0.347	-4.061	-0.003(-0.454, +0.887)
M17	-0.053	-0.000	-0.000				0.399	-3.509	-0.010(-0.543, +0.575)
M9		-0.000			-0.988		0.314	-3.141	-0.133(-0.336, +0.115)
M13	-0.072	-0.033					0.301	-2.869	+0.037(-0.382, +1.031)
M6		0.000	-0.000	0.000			0.359	-2.408	-0.051(-0.514, +0.370)
M14	-0.062	-0.035	0.000				0.349	-2.150	+0.045(-0.398, +0.812)
M8		0.096			-1.253		0.263	-2.093	-0.089(-0.272, +0.111)
M4	-0.067		0.000	0.000		-0.000	0.330	-1.596	+0.354(-0.026, +1.244)
M3		-0.000					0.167	-1.425	-0.022(-0.204, +0.432)
M11					-0.951		0.136	-0.895	-0.046(-0.217, +0.099)
M10			-0.000				0.131	-0.770	+0.015(-0.217, +0.285)
M12			-0.000	-0.000	-0.794		0.222	-0.682	+0.027(-0.238, +0.283)
M2		-0.057					0.050	+1.026	+0.054(-0.104, +0.279)
M5		-0.066	-0.000	-0.000		0.000	0.216	+2.281	+0.113(-0.341, +0.364)

Table 6.2: AICc, mean cross validated ignorance scores (with 90 percent resampling intervals) and R² for each combination of variables ranked in order of AICc. Parameter values are shown where applicable. JF temp, JA temp and GDD correspond to the mean January/February temperature, mean July/August temperature and summer growing degree days respectively. DD and Hunted indicate whether density dependence and the proportion of animals hunted are included as variable and DD Int. Indicates whether the interaction between the climate variable and density dependence is included.

	JF	JA	GDD	DD	Hunted	DD Int	R ²	AICc	Mean Ign
M7	-0.121				-1.069		0.627	-14.511	-0.588(-1.051, -0.114)
M15	-0.112	0.058			1.258		0.671	-14.453	-0.600(-1.101, -0.015)
M18	-0.111		-0.000		-1.071		0.640	-12.974	-0.508(-0.931, +0.015)
M1	-0.120						0.465	-9.936	-0.381(-0.783, +0.150)
M16	-0.110		-0.000				0.478	-8.190	-0.291(-0.768, +0.254)
M13	-0.117	-0.018					0.470	-8.099	-0.273(-0.770, +0.250)
M17	-0.100		-0.000	-0.000			0.492	-6.404	-0.200(-0.728, +0.283)
M14	-0.109	-0.020		-0.000			0.481	-6.218	-0.219(-0.725, +0.265)
M4	-0.102			-0.000			0.475	-5.769	-0.032(-0.381, +0.430)
M9			-0.000		-1.045	-0.000	0.322	-3.233	-0.145(-0.341, +0.107)
M8		-0.099			-1.355		0.284	-2.114	-0.106(-0.294, +0.150)
M6			0.000	-0.000		0.000	0.359	-2.025	-0.106(-0.585, +0.262)
M3			-0.000				0.167	-2.014	-0.045(-0.294, +0.385)
M11					-1.025		0.136	-1.399	-0.039(-0.222, +0.145)
M10			-0.000				0.131	-0.718	+0.004(-0.254, +0.292)
M12				-0.000	-0.856		0.230	-0.676	-0.016(-0.304, +0.252)
M2		-0.057					0.050	+1.391	+0.058(-0.090, +0.312)
M5		0.066		-0.000		0.000	0.216	+1.703	+0.136(-0.302, +0.437)

Table 6.3: The same as table 6.2 but with population change from 2009 to 2010 left out of the analysis

7. Conclusion and Summary of Findings

7.1. Conclusions on Population Modelling

Chapters two to four of this deliverable have focused on methodology for population modelling which has been demonstrated both in the context of the ibex population of Gran Paradiso National Park and the wild reindeer population of Hardangervidda National Park. The use of benchmark models to help determine the information in a set of forecasts has been demonstrated along with a new ‘sanity-check’ test that can be used to estimate the probability that the model selection statistics of a set of models occurred by random chance. Both of these were demonstrated in the context of the ibex population modelling example. To some extent, benchmarking models is something already used in population modelling in that the performance of the ‘best’ model is sometimes compared with that of a ‘null’ model. The work of chapter two, however, expands on this and considers the use of alternative benchmark models that have been found to outperform the ‘null’ model in the ibex population modelling example. The sanity-check test goes some way towards answering the question of how many models should be tested and, to some extent, discourages a modeller from finding a ‘good’ model via brute force (i.e. testing a large number of models and eventually finding a ‘good’ one). Chapter four is then a more methodologically focused version of a paper studying the wild reindeer population of Hardangervidda National Park. The methods demonstrated in the previous two chapters are used here whilst other methodology is considered along the way. This example provides a somewhat different scenario to that of the ibex population modelling example because, at first glance, it is less clear whether the models presented really do contain useful information (rather than appearing to via random chance). The benchmarking approach and sanity-check test introduced in chapters three and four, however, helped to show that there really is information in at least some of the models.

7.2. Some Reflections on Modelling

In both the ibex and reindeer population modelling examples, a number of models were defined a priori and both the relative and absolute performance of each one was assessed. For the ibex example, strong evidence was found to suggest that the best models outperform the best benchmark model and, in addition, that the performance of the models was likely not simply due to chance. In the reindeer example, weaker evidence was found supporting the performance of the best model, although once an influential observation corresponding to a particularly cold winter was removed, the evidence was much stronger. There is a question then of how models such as these can plausibly be used for real world decision making. It should be stressed that evidence that the models significantly outperform benchmark models and levels of significance etc, whilst useful, are not enough to say whether the model can reproduce all of the most important processes and, importantly, make informative predictions. It is important, therefore, to acknowledge the weaknesses as well as the strengths of the model. It is of crucial importance to acknowledge that, no matter how sophisticated the models or statistical techniques that are applied, the data places a fundamental limitation on what is possible in terms of understanding. This is not to say that the models are not useful, however. For example, the reindeer population model implies that, over their typical range, high winter temperatures have a negative effect on the population. This conclusion can be used to design new experiments to investigate this further whilst, for lower temperatures, some form of expert judgement may be necessary.

It should be reiterated that, for the reindeer example, the point corresponding to the coldest year was not removed as an outlier or considered not to be important but that, from the data alone, it is simply not possible to determine the effects of such low temperatures. Whilst it may be the case that the linear relationship does indeed stretch this far (and that that year’s change in population was unusual), to assume that this is the case would be ill judged and may lead to poor decision making. Models can be extremely useful tools with which both to aid understanding and to make predictions of the future. It is important, however, to acknowledge and understand the limitations of models. For example, all of the population models considered in this document assume a linear dependence between some variable and the relative population change. It is clearly the case, however, that the linear relationship cannot extend infinitely in both directions. For example, for density dependence, an extremely high population would almost certainly have a very negative effect on a population. A very small population, on the other hand, may also have a negative effect if animals are dispersed too far apart to mate or if the gene pool is too narrow. It is therefore not a case of whether the linear model is appropriate for the entire range of possible scenarios, but of whether it is appropriate for the ranges of interest. This has important implications for making projections into the future. For example, consider a GLM model featuring linear density dependence such as the Ricker Model. To make projections into the future that feature extinction events, it is almost certainly necessary to extrapolate the relationship between the current population size and the relative population change beyond the points used to select the parameters. Similarly, this may happen under various climate change scenarios when the population’s



response to previously unobserved weather conditions (unprecedented warm summers, for example) is assessed.

Bibliography

- [1] R. C. Smith, *Uncertainty quantification: theory, implementation, and applications*. Siam, 2013, vol. 12.
- [2] L. A. Smith and N. Stern, “Uncertainty in science and its role in climate policy,” *Phil. Trans. R. Soc. A*, vol. 369, no. 1956, pp. 4818–4841, 2011.
- [3] J. Berger and L. Smith, “Formulism for uncertainty quantification (in preparation),” *Annual Review of Statistics and Its Application*, 2018.
- [4] E. L. Thompson, E. A. Sienkiewicz, and L. A. Smith, “When (not) to downscale,” *Nature Climate Change (submitted)*, 2018.
- [5] S. Manabe and R. T. Wetherald, “Thermal equilibrium of the atmosphere with a given distribution of relative humidity,” *Journal of the Atmospheric Sciences*, vol. 24, no. 3, pp. 241–259, 1967.
- [6] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. J. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring *et al.*, “Evaluation of climate models. in: climate change 2013: the physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change,” *Climate Change 2013*, vol. 5, pp. 741–866, 2013.
- [7] K. Judd and L. Smith, “Indistinguishable states: I. perfect model scenario,” *Physica D: Nonlinear Phenomena*, vol. 151, no. 2, pp. 125 – 141, 2001. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167278901002251>
- [8] K. Judd and L. A. Smith, “Indistinguishable states ii: The imperfect model scenario,” *Physica D: nonlinear phenomena*, vol. 196, no. 3-4, pp. 224–242, 2004.
- [9] H. Akaike, “A new look at the statistical model identification,” *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, Dec 1974.
- [10] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [11] S. Amari, N. Murata, K.-R. Muller, M. Finke, and H. H. Yang, “Asymptotic statistical theory of overtraining and cross-validation,” *Neural Networks, IEEE Transactions on*, vol. 8, no. 5, pp. 985–996, Sep 1997.
- [12] R. Mac Nally, R. P. Duncan, J. R. Thomson, and J. D. L. Yen, “Model selection using information criteria, but is the “best” model any good?” *Journal of Applied Ecology*, pp. n/a–n/a, 2017. [Online]. Available: <http://dx.doi.org/10.1111/1365-2664.13060>
- [13] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Apr. 1986.
- [14] E. Pollard, K. H. Lakhani, and P. Rothery, “The detection of densitydependence from a series of annual censuses,” *Ecology*, vol. 68, no. 6, pp. 2046–2055, 1987.
- [15] M. G. Bulmer, “The statistical analysis of density dependence,” *Biometrics*, vol. 31, no. 4, pp. 901–911, 1975. [Online]. Available: <http://www.jstor.org/stable/2529815>
- [16] D. R. Fox and J. Ridsdill-Smith, “Tests for density dependence revisited,” *Oecologia*, vol. 103, no. 4, pp. 435–443, 1995.
- [17] C. J. DiCiccio and J. P. Romano, “Robust permutation tests for correlation and regression coefficients,” *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 1211–1220, 2017.
- [18] R. D. Cook, “Detection of influential observation in linear regression,” *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977. [Online]. Available: <https://doi.org/10.1080/00401706.1977.10489493>
- [19] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.
- [20] C. Bergmeir, R. J. Hyndman, and B. Koo, “A note on the validity of cross-validation for evaluating autoregressive time series prediction,” *Computational Statistics and Data Analysis*, vol. 120, pp. 70 – 83, 2018.
- [21] B. Gompertz, “On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies,” *Philosophical Transactions of the Royal Society of London Series I*, vol. 115, pp. 513–583, 1825.
- [22] W. E. Ricker, “Stock and recruitment,” *Journal of the Fisheries Research Board of Canada*, vol. 11, no. 5, pp. 559–623, 1954. [Online]. Available: <https://doi.org/10.1139/f54-039>
- [23] B. Dennis and M. L. Taper, “Density dependence in time series observations of natural populations: Estimation and testing,” *Ecological Monographs*, vol. 64, no. 2, pp. 205–224, 1994. [Online]. Available: <http://www.jstor.org/stable/2937041>

-
- [24] A. R. Jacobson, A. Provenzale, A. von Hardenberg, B. Bassano, and M. Festa-Bianchet, “Climate forcing and density dependence in a mountain ungulate population,” *Ecology*, vol. 85, no. 6, pp. 1598–1610, 2004. [Online]. Available: <http://dx.doi.org/10.1890/02-0753>
- [25] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2008.
- [26] I. J. Good, “Rational decisions,” *Journal of the Royal Statistical Society: Series B*, vol. 14, pp. 107–114, 1952.
- [27] M. S. Roulston and L. A. Smith, “Evaluating probabilistic forecasts using information theory,” *Monthly Weather Review*, vol. 130, pp. 1653–1660, 2002.
- [28] J. Bröcker and L. Smith, “Scoring probabilistic forecasts: the importance of being proper,” *Tellus A*, vol. 22, no. 2, 2007.

A. Model Selection for Correlated Time Series

Model selection forms a key part of many studies where, typically, information criteria and/or cross-validation approaches are used to attempt to select one or several models with which to make forecasts of the future or to give insights about some underlying process or phenomenon. Often, however, the outcomes form a correlated time series and it is sometimes suggested that, when this is the case, cross-validation can be problematic [20]. Here, however, it is argued that whilst this can indeed be the case, information criteria, often considered a more appropriate approach, can also be problematic and that, often, the limitation is not in the model selection methodology itself, but the data. This chapter does not aim to give a comprehensive overview of the effects of correlated outcomes on model selection techniques. The aim, rather, is to acknowledge some of the problems with cross-validation, discuss the impact of correlation on information criteria and to take a position on the subject of model selection in such scenarios. This chapter can therefore be considered a starting point to LSE's work on this topic.

In a point forecasting context, it has been shown that problems can arise with cross-validation when forecast errors in the training and test sets are correlated because the test set is not representative of a truly out of sample case. It is not necessarily true, however, that correlated outcomes lead to correlated errors [20] and, conversely, it is not necessarily true that uncorrelated outcomes should lead to uncorrelated errors and thus the issue is somewhat deeper than it may first appear. In this deliverable, we are interested not in point forecasts but in probabilistic forecasts. In probabilistic forecasting, interest is shifted from forecast error to, in our case, the logarithm of the probability density placed on each outcome by the forecast (this is the case for the ignorance score, but also in calculating information criteria since the log-likelihood is required). We are thus not interested in the issue of correlated errors, but in whether the log of the densities placed on consecutive outcomes are correlated. Ignoring the issue of parameter selection for now, it is easy to see how these may be correlated. Consider a case in which the mean of a set of probabilistic forecasts is constant over time but that the variance varies in the form of a sine wave. Assume that the outcomes are random draws from their forecast distributions (i.e. the forecasts are perfect). The variance of consecutive forecasts will thus be more similar than two randomly selected forecasts and thus the density placed on the outcome in each case will be correlated (narrower distributions will typically place more density on the outcome). As is widely known in statistics, correlated data tend to lead to longer convergence times than if those data were independent. If the ignorance scores are autocorrelated, it will thus take a larger sample size for the mean ignorance in the sample to converge to the population mean (i.e. that which would be obtained by letting the sample size tend to infinity). For a small sample size, a disproportionate number of forecasts of a certain nature (e.g. with small or large variance in the example described above) can cause inaccurate estimates of the underlying distribution and thus the forecast evaluation statistic may be over or understated. In addition, the parameters may be skewed towards values that tend to exaggerate the performance of over-represented types of forecast in the data.

Correlated forecast performance can be a problem when we are interested in evaluating and comparing forecasts of time series. When this is the case, there is simply less information about out-of-sample forecast performance than if those forecasts were independent. Consider, for example, a case in which two competing models form probabilistic daily maximum temperature forecasts for a particular location and that the aim is to choose the most effective given some evaluation criteria. Suppose that, to do this, forecasts and outcomes from 30 different days are available. In the first case, the forecasts and outcomes are drawn randomly over a course of five years whilst, in the second, forecasts and outcomes from 30 consecutive days are chosen. It stands to reason that the former case provides more information about the general out-of-sample performance of the models since it provides a more representative sample over a year. The latter case, however, represents a shortcoming in the information itself and no methodology can overcome this. It is therefore important to distinguish limitations in the information available from limitations in the methodology. The real question here is, given the limitations in the data, how can we make the best decision regarding model selection?

Perhaps less well documented are the problems correlated forecasts can cause when calculating information criteria. To calculate the AIC, BIC or AICc (among other less well known criteria), it is first necessary to calculate the log-likelihood, which is then weighed up against the number of parameters. If forecasts of a certain kind are over-represented, the log-likelihood (calculated by multiplying the densities placed on each outcome) will be different to that which would be expected were the forecasts independent. The same problems will therefore arise as arise with cross-validation with ignorance. Cross-validation and information criteria differ in how the parameters are selected since, in the former case, a subset of the dataset is used to select the parameters whilst, in the latter, the entire data set is used. In both cases, however, if the forecasts are correlated, the parameter values will be skewed towards those types of forecast that are over-represented. When these are used out of sample, they will provide worse performance, on average, than if the parameters had been estimated using independent data.

To illustrate difficulties in using information criteria for forecasts with correlated performance further, consider the following artificial but illustrative case. A set of n past independent data points, each consisting of outcomes and corresponding explanatory variables, are used to fit two different competing models and information criteria are calculated for

each one. Suppose that model one has a maximised log-likelihood of 21 and 6 free parameters whilst model two has a maximised log-likelihood of 17 and 1 parameter. The AIC is calculated using the formula $AIC = -2\log(\hat{L}) + 2K$ where \hat{L} is the maximised likelihood and K is the number of fitted parameters and thus the AIC for model one is $AIC = -2 \times 21 + 2 \times 6 = -30$ and the AIC for model two is $AIC = -2 \times 17 + 2 \times 1 = -32$. Model two is thus favoured over model one since it has a lower AIC. Now, suppose that each of the data points is repeated exactly once such that the sample size is doubled. The AIC for model one is now $AIC = -2 \times 42 + 2 \times 6 = -72$ and, for model two, $AIC = -2 \times 34 + 2 \times 1 = -66$ and thus the order of the models in terms of the AIC has been reversed. The result here would be model overfitting since the model with the largest number of parameters is wrongly selected. Whilst, clearly, it makes no sense to use each of the data points twice, this example exposes an important shortcoming in the use of information criteria since there is not a great deal of difference between repeated and extremely similar data. For example, two weather forecasts made two minutes apart but with the same lead time would be extremely similar and would have extremely similar outcomes. The issue here is, in fact, that the sample size and the 'effective' sample size are different. In the case outlined above, the sample size is $2n$ but effectively n since there is no extra information in the repeated data. The AIC, however, does not account for this and thus can give misleading results such as in the example described above.

All of the above leads to the question of how this affects the population modelling in this deliverable. In most cases, the models considered use weather conditions as well as the current population as inputs whilst the variance of the probabilistic forecasts is fixed. This means that, were the forecasts to be perfect (i.e. the outcome were truly drawn from the forecast distribution), since the variance is not dependent on the inputs to the model, the distribution of densities on the outcome (and consequently, the ignorance score) given a perfect forecast would be constant over all forecasts and thus there is no problem with using AIC or cross-validation. In practice, of course, the forecasts are not perfect and thus it is difficult to say whether model error is state dependent. Since both information criteria and cross-validation are impacted negatively by correlation in forecast performance, however, it is often worthwhile to use both and, if the ordering of models is broadly similar, more confidence can be had in the results.

B. Background Methodology For Population Modelling

In this appendix chapter, frequently occurring background methodology is described that is referred to throughout this document. Any background methodology that is used only once is described in its relevant section.

B.1. Population Models

Population models are models that can be used to attempt to understand the dynamics of a population. These often make simple assumptions about how populations grow and decline and can often be extended to include other important input variables. In this document, modified versions of two well known population models, the Stochastic Ricker and Stochastic Gompertz Models, are used to model wild animal populations. To define these models, denote some population in the i th year to be n_i and define $R_i = \log\left(\frac{n_{i+1}}{n_i}\right)$ to be the *relative population change* from the i th to the i plus first year.

An important concept in population dynamics is that of *density dependence*. A population is said to be density dependent if its population growth or decline depends on the current population. The stochastic Gompertz model [21] is a model of density dependence defined as

$$R_i = a + b \log(n_i) + \sigma \epsilon \quad (\text{B.1})$$

and the Stochastic Ricker Model [22, 23], which assumes a slightly different form of density dependence is defined as

$$R_i = a + b n_i + \sigma \epsilon \quad (\text{B.2})$$

where, in both cases, a and b are parameters to be selected and ϵ is a random draw from a Gaussian distribution with mean zero and standard deviation σ . Both of these models are generalised linear models and thus the values of a , b and σ can be found simultaneously through standard least squares estimation.

Both of the above models can be modified such that extra explanatory variables for the relative population can be added. The modified Gompertz model [24] is defined by

$$R_i = a + b \log(n_i) + \sum_{i=1}^V c_i V_i + \sigma \epsilon \quad (\text{B.3})$$

and the modified Stochastic Ricker model [24] is defined by

$$R_i = a + b n_i + \sum_{i=1}^V c_i V_i + \sigma \epsilon \quad (\text{B.4})$$

where, in both cases, V_i is some explanatory variable, and c_1, \dots, c_V are extra parameters to be selected.

The Modified Stochastic Ricker Model is used extensively in this document both for the modelling of ibex populations in Gran Paradiso National Park and wild reindeer populations in Hardangervidda national park. The Modified Gompertz is used only to model ibex populations.

B.2. Model Selection Techniques

In statistics and other scientific disciplines, it is very common for a practitioner to have a set of different models of the same phenomenon or process. As a result, it is often desirable to choose the most appropriate model based on its performance over some past set of data. This process is called *model selection*. The issue of model selection is one that has been studied in great detail. Most applications of model selection use methodology that falls into two different categories: information criteria and cross validation. Information criteria weigh up the in-sample fit of each model with the number of empirically chosen parameters in that model to attempt to assess how well it would perform out of sample. Cross-validation approaches fit the parameters over a training subset of the data set and test their performance on another distinct subset. The process is then repeated using different training subsets. Both approaches to model selection are now described in more detail.

B.2.1. Information Criteria

Information criteria provide an information theoretic approach to the problem of model selection. In all cases, the ‘fit’ of the model to the data set is weighed up against the number of parameters in order to avoid overfitting. The two most well known information criteria are Akaike’s information criterion (AIC) [9] and the Bayesian information criterion (BIC) [10]. The formulation of the two differ only in how they penalise extra parameters. AIC is given by

$$\text{AIC} = -2 \log(\hat{L}) + 2K \quad (\text{B.5})$$

and BIC is given by

$$\text{BIC} = -2 \log(\hat{L}) + K \log(n) \quad (\text{B.6})$$

where K is the number of parameters selected from the data and n is the sample size. In each case, the model with the lowest AIC or BIC is considered to be the most appropriate when applied out of sample. Although the formulations of AIC and BIC are very similar, they are in fact derived in very different ways. Most notably, BIC is derived from a Bayesian paradigm and makes the assumption that one of the models is the true model. AIC, on the other hand, does not assume that any of the candidate models generated the data and is a consistent estimator of the Kullback Leibler Divergence between the model distribution and the true distribution. For small sample sizes, however, AIC is slightly biased and thus a corrected, unbiased, version is often used. The corrected version AIC_c [25] is defined by

$$\text{AIC}_c = -2 \log(\hat{L}) + \frac{2K(K+1)}{n-K-1}. \quad (\text{B.7})$$

Note that AIC and AIC_c are asymptotically equivalent and so they only differ significantly when the sample size is small. Consistent with the approach taken in the original paper and given the relatively large sample size, AIC is used to compare models for ibex populations. For the wild reindeer example, the sample size is smaller and thus the corrected version AIC_c is used.

B.2.2. Cross Validation

Cross-validation is a method of model-validation used to attempt to assess how well a model will perform out of sample. As such, it can be used to assess which of a variety of models would perform ‘best’ out of sample and it can thus be considered a model selection technique. The key feature of cross-validation is that the data set is divided into a training set, over which the parameters are selected, and a test set, over which the model with those parameters is tested. The process is then repeated with different subsets of the data set used as the training and test sets. One particular form is leave one out cross validation in which the test set consists of a single point and the training set consists of each of the other points. This process is repeated such that each data point forms the test set exactly once.

B.3. Probabilistic Forecast Evaluation

In this paper, probabilistic forecasts are evaluated using a function of the forecast and the outcome called a scoring rule. A scoring rule with useful properties is the *ignorance score* [26, 27] defined by

$$\text{IGN} = -\log_2(p(Y)) \quad (\text{B.8})$$

where $p(Y)$ is the probability density placed on the outcome. The ignorance score is negatively oriented and hence smaller values indicate better forecast skill. An advantage of the ignorance score is in its interpretation. The difference in the mean ignorance between two sets of forecasts can be interpreted as the base 2 logarithm of the ratio of the density placed on the outcome by each, measured in bits. For example, if the mean ignorance of one set of forecasts is 3 bits smaller than another, it places 2^3 times more probability density on the outcome, on average. Another benefit of the ignorance score is that it is *proper* [28]. A proper score is optimised when the distribution from which the outcome is drawn is issued as the forecast. Propriety encourages a forecaster to issue their true belief as the forecast.

B.4. Assessing Model Significance

It is often desirable to test whether the skill of two sets of forecasts are significantly different. A simple way to assess the significance is to use resampling. When the forecasts are assessed using the ignorance score, this can be done using a simple paired test.

Define the ignorance of a set of forecasts from one forecasting system to be $\text{ign}_1^a, \dots, \text{ign}_n^a$ and the corresponding ignorance of the a second forecasting system be $\text{ign}_1^b, \dots, \text{ign}_n^b$. The relative ignorance of the i th forecast is defined by $\text{ign}_i^{\text{rel}} =$

$\text{ign}_1^a - \text{ign}_1^b$. A bootstrapping approach can then be used to find a p-value for the null hypothesis that the mean relative ignorance is zero compared with the alternative hypothesis that one forecasting system outperforms the other. s

B.5. Ibex in Gran Paradiso

In this document, a population study of ibex is used as an example on which to demonstrate the methods presented. The data and models come from the paper ‘Climate forcing and density dependence in a mountain ungulate population’ [24]. This paper focuses on the modelling and future projection of the population of ibex in Gran Paradiso national park in northwestern Italy. An introduction to this study is given here whilst more details are given as and when required throughout the document. The paper attempts to identify variables that drive changes in the size of the population. These variables are used as inputs to the modified stochastic Ricker and Gompertz population models and the performance is compared using Akaike’s Information Criterion (AIC). The model that is found to have the smallest AIC (i.e. the best fitting model) is used to project the population (given already known weather conditions) over an extended period of time (20 years) and is shown to produce similar qualitative behaviour to a population explosion seen in the 1980s.

B.5.1. Data

The paper considers counts of the ibex population taken annually from 1956 to 2000. This time series is shown in figure B.1. This data set is considered of high importance in ecology because it is the longest data set of its kind in the world. In addition to population estimates, the paper considers a number of climatic variables derived from two nearby weather stations. These variables consist of daily observations of minimum and maximum temperatures, precipitation and snow depth. These were available from 1959 and 1962 respectively for each weather station. The observations were aggregated to create indicators of the climate in each year. Both the modified Stochastic Gompertz and Ricker Models with various combinations of variables were considered as candidate models. Different combinations of the current population, the mean snow depth and the interaction between the two were considered for each. The AIC for each model was calculated and the models ranked according to this measure. Details of each of these models are shown in section 4.4.

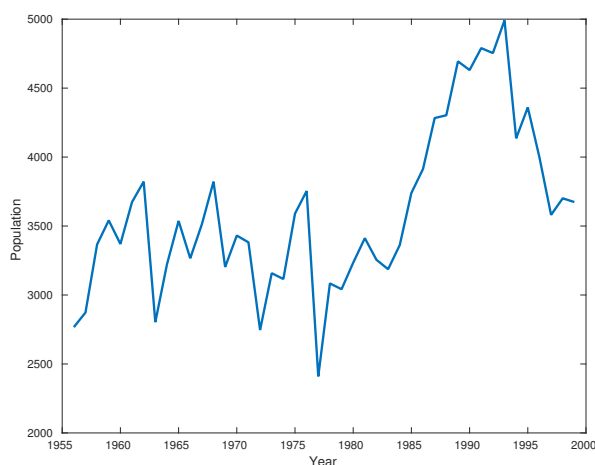


Figure B.1: Time series of ibex population numbers in Gran Paradiso National Park between 1956 and 1999.

C. Understanding Uncertainty in Environmental Modelling

C.1. Workshop description, aims and objectives

The workshop was held from 12-14 September 2017 at the Royal Society in London, organised and hosted by Dr Erica Thompson. Expert lecturers included Dr Erica Thompson (course organiser), Professor Leonard Smith, Dr Ed Wheatcroft, Dr David Stainforth (all from LSE's Centre for the Analysis of Time Series) and invited speaker Dr Elisabeth Stephens (Reading University).

The course was advertised as follows: *Modelling and simulation are an increasingly important part of modern science, especially in highly policy-relevant disciplines such as ecosystem modelling, natural hazards, and climate projection. Good practice in the use and interpretation of models is therefore vital, both for sound science and for informing evidence-based policy decisions. The workshop will present an overview of model evaluation methods, statistical inference for model output, and the use of models in risk management and decision-making, with the aim of exposing participants to methods and insights available in environmental modelling and encouraging critical evaluation of the approaches and methodologies used in their own research. The workshop will be structured around several themes, with facilitated discussion time and interactive problem-solving exercises, allowing participants to explore and understand the concepts presented by the expert lecturers. Participants will leave with an overview of the key issues of uncertainty in environmental modelling, an understanding of how these affect their own research methods and where to find expert guidance and further information. The course includes a poster session and practical exercises and is held this year in the inspiring surroundings of the Royal Society, in London.*

The main aims of the workshop were:

- to communicate some key results about the use and interpretation of environmental models;
- to discuss these themes with a broad range of early-career researchers;
- to help participants apply the insights gained to their own research projects;
- having lifted the lid on many important themes over just three days, to signpost where to find further information and insight;
- to create a community of early-career researchers who are aware of the challenges of model use and interpretation and are still at a point where they can take account of these in their research careers.

C.2. Workshop content

The first half of day 1 was spent ensuring that all participants were familiar with the terminology and basic concepts that would be used throughout the course, briefly introducing everyone to each other and starting to think about how their own use of models fitted into the bigger picture.

Guest lecturer **Dr Elisabeth Stephens from Reading University** presented a talk on communication and use of forecasts, with a focus on real-world decision-making and risk management. This set the scene for a continued emphasis throughout the workshop on making forecasts useful and accountable to the end user decision-maker, both by appropriate design of the model itself and by taking care with communication.

Dr Ed Wheatcroft then took a closer look at the details of quantitative forecast evaluation, considering some methods and metrics that may be of use to participants. This lecture laid the basis for many things that we would discuss again later in the course. The following topics were covered with a basic explanation and signposts to further material:

- Differences between point and probabilistic forecasts.
- The case for probabilistic forecasting.
- Evaluating probabilistic forecasts.
- Focus on binary forecasts: forecasting the probability of scoring in football.
- Reliability Diagrams as diagnostic tools.

Public forecasts of the probability of scoring in football were used as an accessible example of how to make and evaluate probabilistic forecasts, and how to use the reliability diagram as a diagnostic tool for further evaluation and development.

Professor Leonard Smith then ran an interactive lecture structured around a series of games (with prizes for the winners to incentivise full participation). These illustrated various concepts of real-world decision-making, including the pitfalls of making “rational” decisions based on a model which is inadequate (supplies misleading probability forecasts) and what sorts of alternative strategies can be used when this is the case.

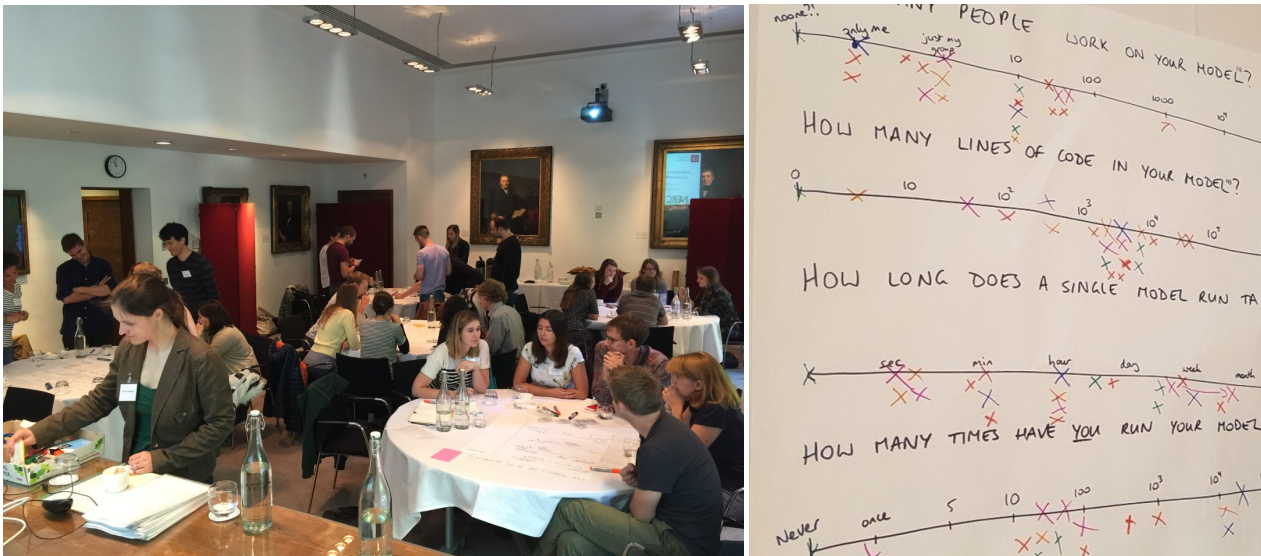


Figure C.1: The atmosphere of the workshop, as intended, was informal and collaborative with small group discussions ensuring that all participants had a chance to share their views and experiences on the material covered.

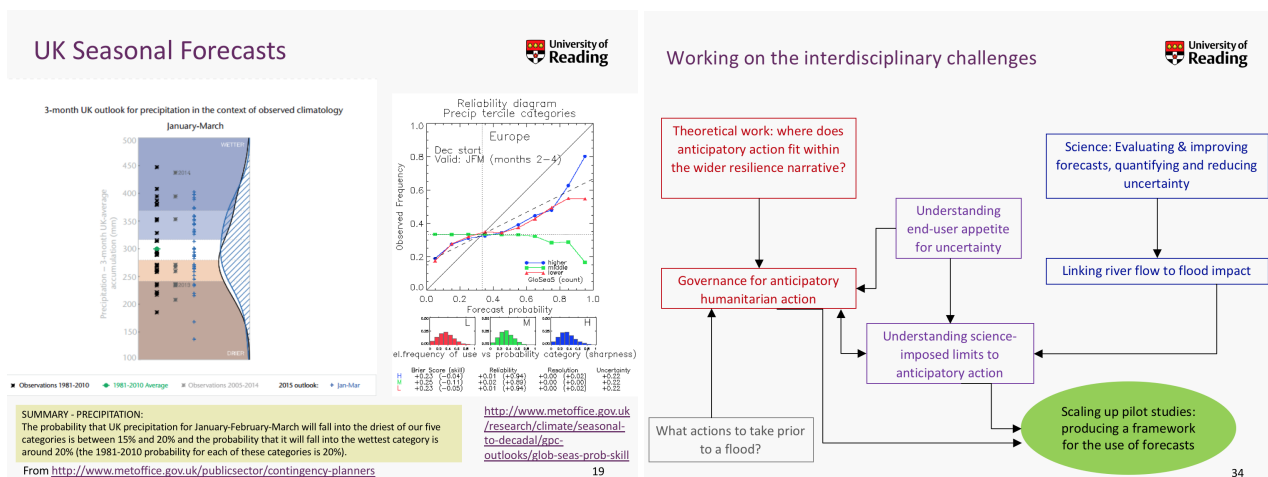


Figure C.2: Dr Stephens' lecture was an entertaining and interactive introduction to the challenges of real-time forecasting for real-world decision-makers.

Later, a **research paper activity in small groups** gave participants an opportunity to read and closely examine a choice of papers using different statistical methods for processing/interpreting model output. The options were:

- statistical emulation;
- Generalised Likelihood Uncertainty Estimation (GLUE);
- use and interpretation of ensembles of models;
- Bayesian history matching;
- model risk management; or
- principles of uncertainty quantification.

Each group had a paper to read and a set of questions to answer collaboratively, based on the concepts presented in the paper but encouraging wider thinking about the theoretical basis and applicability of the methods. The participants enjoyed going into more detail and finding out about specific methods that they either were already using or may be able to use in their research projects.

Next, **Dr Erica Thompson** gave a talk considering the difference between “weather-like” situations – where there is a large enough archive of forecast and outcome (model and reality) pairs that we can undertake a reliable quantitative evaluation of accuracy – and “climate-like” situations, where we are in an extrapolatory regime and do not necessarily expect past performance always to be a reliable guide to future success. Some of the methodological background was explored in more detail, with an introduction to in- and out-of-sample data, the reference class problem, model selection, sampling biases, cross-validation, epistemic uncertainty and the Hawkmoth Effect.

Heat Map of Forecast Values

- The forecast probability of each shot directly resulting in a goal is shown in the plot below.
- Much higher probabilities are placed on shots closer to the goal.
- Shots taken from near the penalty spot are much more likely to result in goals (because these are usually from penalties!).

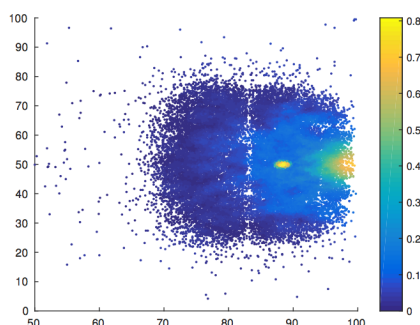


Figure C.3: Dr Wheatcroft presented a comprehensive introduction to the principle and practice of quantitative forecast evaluation, including examples from sporting events.

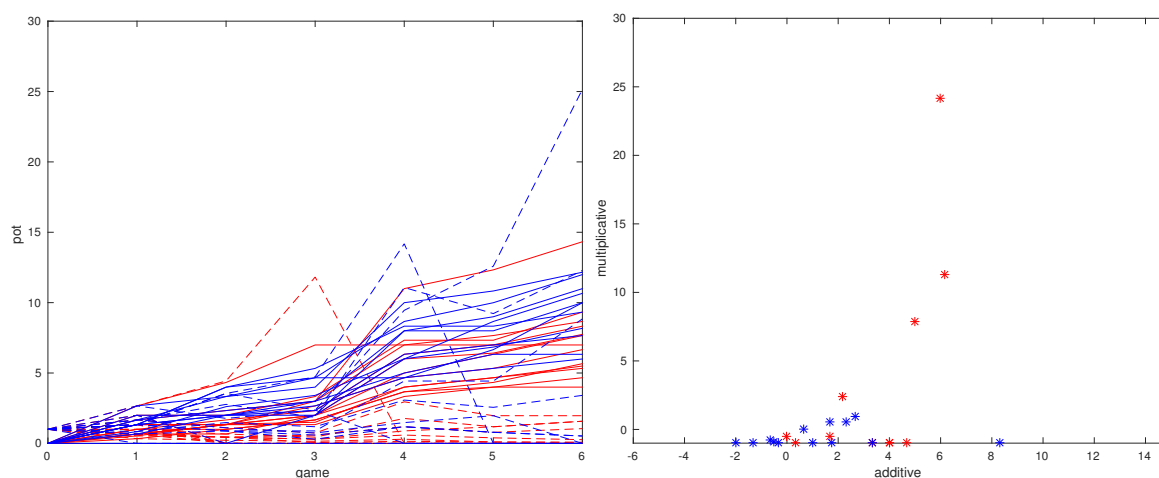


Figure C.4: Left: Cumulative profit of each player over 6 games given additive (solid line) and multiplicative (dashed line) profit/loss coloured by whether the player opted to cut their hands off (red) or not (blue). Right: Scatter plot of total profit given additive (x axis) and multiplicative (y axis) profit/loss, coloured in the same way.

The Butterfly Effect is well-known as the sensitivity to initial conditions displayed by some dynamical systems, meaning that a small perturbation to initial conditions can result in a large change to the state of the system after some length of time (dynamical instability). The Hawkmoth Effect, by analogy, is the sensitivity to structural model formulation, meaning that a small perturbation to the system itself can result in a large change to the state of the system after some length of time (structural instability). We discussed some of the potential consequences for model construction, calibration and evaluation.

Following from the talk by Dr Thompson, we then considered a detailed practical example of the implications of limited data, model inadequacy and the Hawkmoth Effect. **Dr David Stainforth** discussed the use of climate model ensembles with reference to some recent papers (including one studied in depth by a group in the previous activity). He illustrated a number of important statistical challenges for the interpretation of model ensembles, including:

- lack of independence;
- difficulties with model weighting by performance (including culling);

Limited data

- When does this happen?
 1. System exists but is poorly observed
 Difficult to get to: deep ocean sediments
 In the past: palaeoclimate, fossil record
 No funding or obs have only recently begun
 Not clear what is relevant: much of economics/social sciences
 Data are **applicable but not available**
 2. We are interested in out-of-sample behaviour
 - o out-of-sample: Future climate
 - x out-of-sample: Exoplanets
 - θ out-of-sample: Biscuit preferences in New Zealand
 Data are **available but not applicable**



1. Applicable data, but not much of it

- What are the key challenges?
 - Calibration of model and evaluation of model must happen separately
 - Beware of double-counting
 - How representative are the data?
 - Beware of sampling biases
 - Attempt to quantify possible sampling uncertainty
 - Many models may fit the same data
 - Don't leap to select "best" model based on data, even if the fit is very good
 - Precious data: use it carefully

Figure C.5: Dr Thompson introduced some concepts around methodology in situations of limited data (complementary to Dr Wheatcroft's presentation about how to proceed when plenty of data are available for calibration and evaluation).

The Hawkmoth Effect

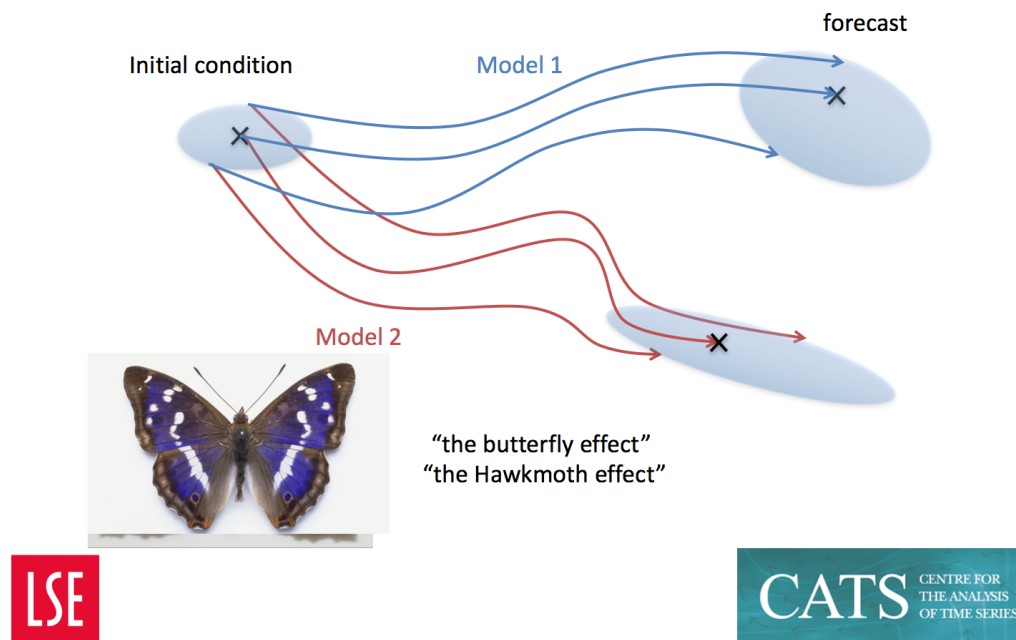


Figure C.6: The Hawkmoth Effect is an effect of structural instability, by analogy with the Butterfly Effect which is an effect of dynamical instability. In essence, the Hawkmoth Effect tells us that for a complex system, we could be arbitrarily close to having the "correct" dynamical equations (the correct physics of the situation) but still get arbitrarily wrong answers.

- the problem of in-sample analysis;
- epistemic issues with statistics generated over different "possible" model versions.

The conclusion of the workshop was a **keynote lecture from Professor Leonard Smith** on *How to do good science with mathematical models*. This was an ECOPOTENTIAL lecture and the transcript and slides are in Appendix D.

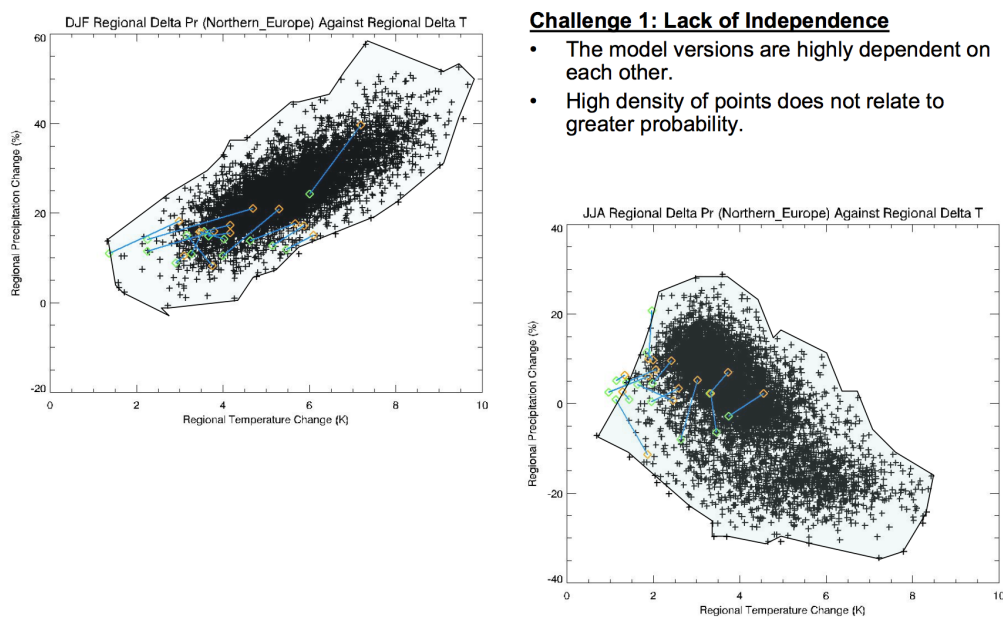
We also gathered some **feedback from participants**:

"Very good course to improve my understanding of uncertainty, and in the future I will aim to be more transparent and to quantify my uncertainties in a more robust way"

"Really opened my eyes to sources of uncertainty I had not considered! Will greatly improve my critical approach to models."

"I feel better equipped to evaluate models and stat outputs both in papers I read and the models I use"

Regional Distributions



Challenge 1: Lack of Independence

- The model versions are highly dependent on each other.
- High density of points does not relate to greater probability.

Figure C.7: In his talk, Dr Stainforth referred to a number of challenges for the interpretation of (climate) model ensembles, the first being lack of independence and the consequent difficulty of generating decision-relevant probabilities from ensemble density in any metric.

“It helps me to step back to think about what I really want to achieve with modelling and whether the uncertainty that I am going to report is meaningful/useful at all”

D. UUEM Lecture Slides and Transcript

Title: How to do good science with mathematical models

Speaker: Professor Leonard Smith

Date: 14 September 2017

**Event: Understanding Uncertainty in Environmental Modelling Workshop, 12-14 September 2017,
The Royal Society**

LSE CATS CENTRE FOR THE ANALYSIS OF TIME SERIES
www.lsecats.ac.uk
www.cccep.ac.uk

14.00 Lecture: How to do good science with mathematical models
Prof Leonard Smith, LSE and University of Oxford

Leonard A. Smith
London School of Economics
&
Pembroke College, Oxford

Logos: CRUI SSE, G L I M P S E, ECO, and others.

Introduction by Erica Thompson

Alright, let's get started again. No doubt you are on tenterhooks to find out the result of the competition yesterday and it will be part of Lenny's discussion today, hopefully.

Leonard Smith

This is a record! I have never actually constructed my talk while it was already started.

Those two are just to remind me to come back at the end to talk about taking a position. While I have had people make slides for my talk while it had already started. That is what Ed is doing now. You can send me the NFL one as well, the sheet from last week.

Ok, did you have a good morning? Did you have a good evening?

It was a placeholder. That was the same way that Ed Lorenz got the whole butterfly moniker and the butterfly effect, which led to the hawkmoth effect. Because Ed didn't send a title in in time. And so the placeholder title became the title. Actually, the original butterfly was a seagull. It was a seagull in Brazil, not a butterfly in Texas.

"How to do good science with mathematical models." Why was this a good placeholder? I think there is a real difference between good science and good maths. You can't really be doing good science and bad maths, but you could be doing good maths and bad science. And you could also be doing bad maths and bad science. But that one is not really what we are going to talk about.

Focus on YOUR question!

**Your actual question, neither
the maths, nor the models.**

The tools are not the product.



?Cambridge or NASA?

LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

So the first thing I would suggest is that *you* really focus on *your* question, on your real question not the thing the grant says (except for people at LSE) not the thing the people you are working with say, Your actual question, not the maths not the models. The tools really are not the product. That is the thing - the goal, unless you are a mathematician, the goal isn't a mathematical product.

For my first post doc, I had the option of staying with NASA in New York, or going to Cambridge. My supervisor more or less said, "Look Lenny, if you go to Cambridge you might end up in England for 25 years, but if you stay at NASA you can do everything they want on Monday, and you put it in your drawer, and then you do what you want on their computers (which are bigger than the ones in Cambridge), and then when they come in you pull open the drawer and show them what they wanted, and they will go away, and they will be very happy. So in that case you are finding ways to facilitate what you are trying to do. The problem comes, or the challenge comes, if you end up really focussing on the tools and not the question.

Fallacy of Misplaced Concreteness

“The advantage of confining attention to a definite group of abstractions, is that you confine your thoughts to clear-cut definite things, with clear-cut definite relations. ... The disadvantage of exclusive attention to a group of abstractions, however well-founded, is that, by the nature of the case, you have abstracted from the remainder of things.

... it is of the utmost importance to be vigilant in critically revising your *modes* of abstraction.

Sometimes it happens that the service rendered by philosophy is entirely obscured by the astonishing success of a scheme of abstractions in expressing the dominant interests of an epoch.”

A N Whitehead. *Science and the Modern World*. Pg 58/9



Whitehead was criticising the straightjacket of Newtonian science; today, perhaps, computer simulation may impede more than just the progress of science.

In the real world, mathematics is never rigorously relevant (beyond the integers!)

This is Alfred North Whitehead and he is talking about the advantages of doing abstractions as ‘modes of thought’. If you find this I suggest you read these pages. If you substitute Modes of Thought, (he was writing in the 20s) for models, computational models (which didn’t exist in the 20s), everything sort of more or less holds. What he is warning people about was actually Newtonian dynamics, Newtonian science, the Newtonian way of thinking about how a problem was solved. That was a straightjacket that had held back science for a hundred years. I would argue that computer simulation is somehow taking that same role today. If you can’t simulate it, if you can’t make pretty pictures if you can’t do lots and lots of statistics that actually have no foundation in reality, then you really are impeding the progress of science. I would also argue that in the real world (what Whitehead is arguing, both as a philosopher and a mathematician) first rank, mathematics is never rigorously relevant. It is just not, well maybe for the integers, but not as soon as you are worried about real numbers your fundamental questions. If you are on a digital computer you are not worried about real numbers, but you still have really fundamental questions.

Would you rather be:

Rigorous

OR

Relevant

TSA

ATS

?

CATS CENTRE FOR THE ANALYSIS OF TIME SERIES 13 Sept 2017 Predictability Probability and JEDI Insight University of Reading Leonard Smith

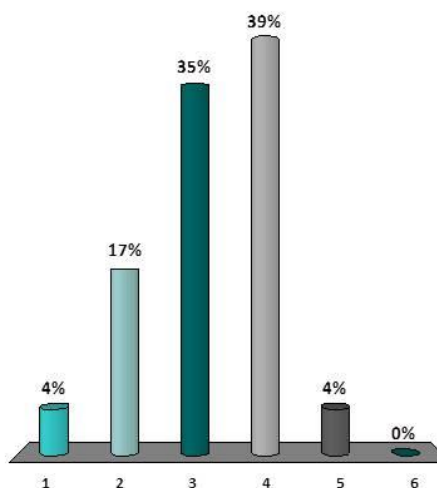
So you have to make a decision about whether you really want to be *rigorous* or if you really want to be *relevant*. And this again this is a decision that mathematicians have to be rigorous, and physicists and statisticians and environmental scientists can pretty much decide when they want to be which if you are over here (rigorous) you are doing time series analysis. At CATS we do the analysis of time series, so this is where the 'ATS' of CATS comes from. We have just had a paper published to appear in the SIAM Journal of Uncertainty Quantification. In the first referee report - this is with Du - the first referee report noted that the word 'non-rigorous' appeared *seventeen times*. And they were right because: we made assumption where to be rigorous we would have had to define what the system actually was, right? So we have the real world that is the system and we have the model that we know mathematically, and we have the computer models, which we sort of know mathematically. I cannot assume I know the system or I am over here in mathematics; so you cannot be rigorously relevant in a mathematical sense and still do the real world case.

If you want to come over here (rigorous) you are going to have to make some assumptions that are indefensible and there are huge numbers of people who try to defend them. I would suggest you not waste your time, you just accept this as a limitation of what we do. The editors, maybe even the referee in this case, did accept that. Eventually.

What else do you need to accept? Well, how many of you started off with your priors on the coin spinning about fifty-fifty or sixty-forty? Right? I sort of knew when I first learned this that wasn't the right answer but I didn't know which side to go for. This is what happened.

The number of times my coin stopped heads up:

1. Five
2. Four
3. Three
4. Two
5. One
6. Zero



There are really cool questions to ask about well, are these multiple models how much did you learn about your coin or does your coin actually have a probability if you spun it, once you spun it 1024 times. Maybe that probability has changed, or maybe the probability is empirically vacuous. (And there is just no such thing as that probability). But it is still a useful tool. It is still a useful thing to talk about even if it does not exist (sort of like the macroscopic 'laws of physics').

Focus on YOUR question!

The tools are not the product.

Observations Rule



LSE UUEM

Royal Society

Thur PM

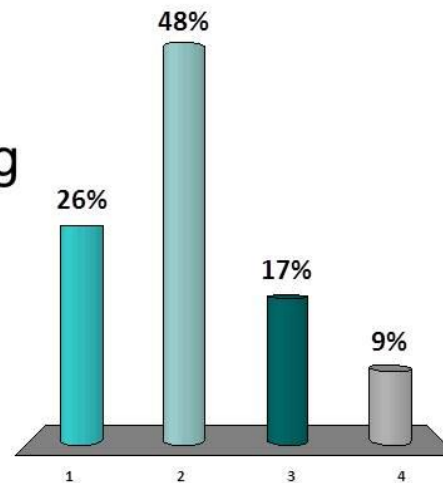
14 Sept 2017

Leonard Smith

But once you see this, right, this is the real world, whichever bin your coin fell in or however much you think you might have been in the tails of the distribution only by chance: **the observations rule**, so my second point: the tools are not the product. Observations rule. If you are a mathematician the observations just don't matter; you make another assumption. But that is all you have to do. The key is that your results flow from your assumptions. You just keep making assumptions until you can make the proof. That is all you need. But once we see the real world is not symmetric we don't need to do a statistical test even to say this isn't symmetric. So then how are we going to deal with whatever the question is if you were actually in the process of spinning coins?

My primary interest/challenge/Question(s) is:

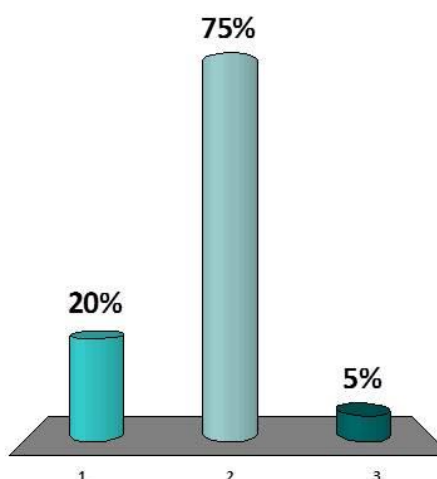
1. Weather-like
2. Climate-like
3. Sub-Model Modelling
4. Elsewhere



So what you said yesterday about your interests. At the moment, I am interested in looking at weather-like. I don't know what David said earlier about climate-like, but this is a really cool place to be and it is cool for a couple of reasons. I mean, it is really hard but there is a great deal of epistemological structure that just hasn't been built for this kind of problem, or *this* kind of problem.

My System of interest is:

1. Weather-like
2. Climate-like
3. Other



**Nothing to do with forecasting here;
the issue is one of evaluation/verification.**



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

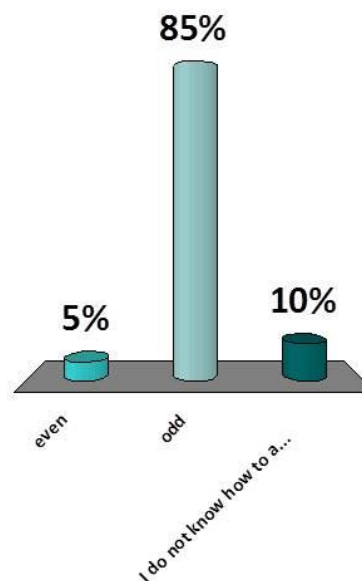
Leonard Smith

This is also really cool. So thank you very much, this was not one of my categories before the things you said yesterday. I now have a new category to think about. But even this one, the foundations here are still under construction and that is a really, really cool time to be in a science.

So Charney, Charney is a very, very famous, rightly so, atmospheric guy, worked with van Neumann on the very first weather/atmosphere models. He wasn't sure whether to go into some very well developed field or meteorology and he went to a very famous guy and said "What should I do?" something in aeronautics, the guy basically said no, meteorology is really young now, that is the place to be, and this was in the '40s. In a lot of these kinds of problems, however hard they may sound, even merely developing the ideas of to approach these problems is still open. So don't worry if people tell you that one can't do things that have been standard for fifty years; you are not alone. This was yesterday afternoon at the University of Reading. This is out of about 36 new students for their graduate programme - this cool graduate programme of maths and physics and meteorology and climate and everything else.

Clickers: Is today's date even or odd?

- A. even
- B. odd
- C. I do not know how to answer this question



Ok so this was... I originally thought this was going to be a clicker test question to make sure, to see how many people will answer. In fact it is not just that. I am not sure which one of these was you, but I almost always get somebody over here (the ten per cent). I make this joke about how this was the pure mathematicians in the room and the date was actually on the slide. If you really want to do physics with maths or physical sciences with maths you sort of have to allow people a way out. You know, things that are really obvious to you, but may not be as well defined as the date.

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists)



LSE UUEM

Royal Society


Thur PM

14 Sept 2017

Leonard Smith

So I guess you want to watch out for lobbyists and jokers, but you also want to allow people to be able to express the fact that they are really confused in a non-embarrassing way. And you need to really not forget that you probably know more about your particular problem than anybody else in the room.

I think this is another thing to think about: when you are presenting results, especially when you are presenting results that could affect people's house prices, their lives, the lives of their children, or grandchildren, a great deal of motivation comes from this sort of fear, fear that the cure might be worse than the disease. Fear that you are actually saying something that is too bad to be true. I think I said yesterday this thing about the last total eclipse, that eventually the sun will expand and vaporise the planet. And sometimes I will get more questions after a talk on that statement - they just don't want the sun to ever vaporise the planet and they have missed the rest of my talk. They were just worried about the sun vaporising the planet. So this is a huge motivating factor.



**We are walking in Florida.
You find you have just been bitten on the hand by a snake.
We did not see the snake.
If it was the deadly carbonblack snake, the bite will kill you
in a painful way, unless you cut off your hand within 15 secs.
I have a hatchet.
You have 9 seconds left.**

CATS CENTRE FOR THE ANALYSIS OF TIME SERIES
LSE UUEM Royal Society Thur PM 14 Sept 2017 Leonard Smith

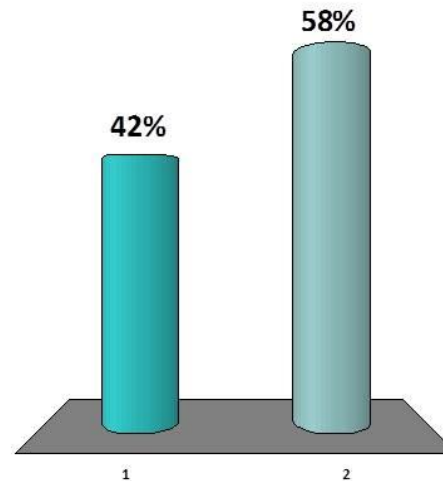
Ok, so the deadly carbonblack snake. These are in fact outlines of pieces of coal. This was not supposed to be a real snake, but I was asked again yesterday how common these snakes were in Florida.

Did you cut off your hand? (in time)

1. Yes
2. No

Note: failing to decide is in fact a decision!

And one does not require full quantitative probabilistic information in order to make a decision.



So did you cut your hand off in time? This was one answer.

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists/Fear**)**


Achievable Scientific Goals

**When I arrived on Capitol Hill I knew
Only one thing about climate change:
Al Gore was in favour of it.
That was enough.**

Bob Inglis

**I suggest you make clear conscious
decision between pursuing
“science to inform” or
“science to motivate”.**

So I really like Bob Inglis now. This statement taught me a lot. When he arrived on Capitol Hill he knew just one thing about climate change – Al Gore was in favour of it and that was enough. And the reason these Republicans were so against... they weren't just against Gore... they weren't just against the Democrats. The idea of the things you would have to do, that they thought you would have to do, this is what was so repugnant to them. They were afraid of them. They thought that literally could be worse than losing half of Florida, every major big city on the eastern seaboard, I mean somehow this idea of mitigating climate change was worse. And if you don't respect that then it will be really hard to make contact with those kinds of people. And this is again why I really want to stay in this 'science to inform'. Once you get in this 'science to motivate' you start becoming blind to some of these issues and then it doesn't matter how good your maths is, or how relevant your models are.

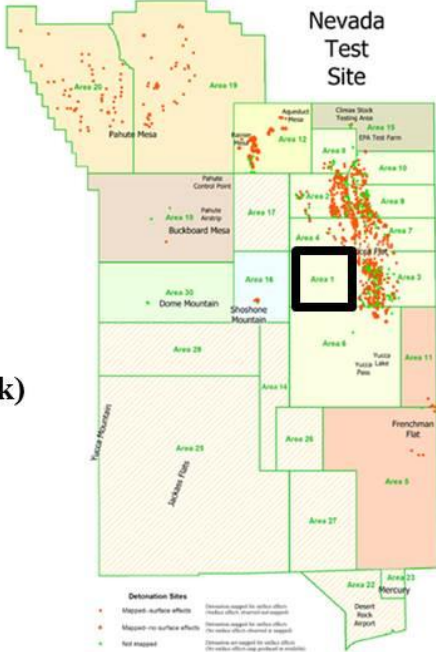


1000 ft Ua1 shaft

Chasing Model Inadequacy

Ball(s)

- 2 bowling balls**
- 3 Basketballs**
- 2 golf balls**
- 3 Wiffle balls**
- ... (no rubber duck)**



Nevada Test Site

Factors

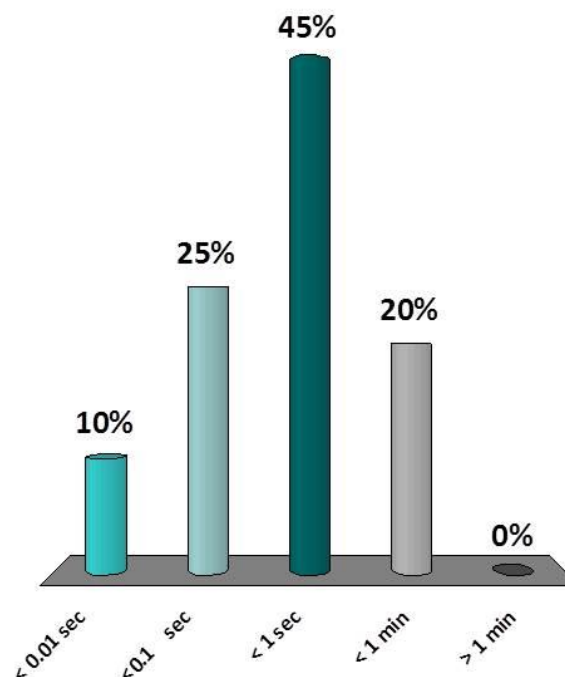
<http://www2.nstec.com/Documents/Fact%20Sheets/U1a%20Facility.pdf>

Thur PM 14 Sept 2017 Leonard Smith

How close was our mean time for basketballs?

- A. < 0.01 sec
- B. < 0.1 sec
- C. < 1 sec
- D. < 1 min
- E. > 1 min

Basket ball.
Initial velocity zero.
1000 ft “tower”.
Laser sheet timing.



Q3.2

So the model inadequacy question. This was you yesterday. I still don't know why you have been better. I would have been but you know we missed that. The answer was here. The point is: if we had thought harder, or better, would someone come up with 'does not reach the bottom'? Why didn't we think of this? So these sorts of questions: what could really go wrong? You are usually in such a rush to finish the talk before you actually get to the place that you are giving it. Just step back (maybe after you get the first review) and think about what the reviewer said. Or try to think about what could really be wrong?

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists)

Think Harder



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

Someone just tweeted... Tamsin just tweeted, that she is going to a meeting to discuss the worst possible outcome for sea level rise in the next hundred years. That is pretty easy isn't it? All land ice slides into the ocean, but that is a physically plausible it must be more probable than an asteroid hits the earth, mustn't it? Well Erica doubts it. Well maybe it is not even more, but maybe we should be worried about an asteroid hitting the earth and vaporising the ocean and sea level goes to zero. Either way, I think there are lots of times when it is really actually quite useful to think out in the physically conceivable, but no that couldn't possibly happen. The ball doesn't make it to the bottom. If someone had said that then how could that happen and would have had somehow the ball hits the wall and we would have put some probability there. So think harder.

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

AR4 Summary for Policymakers

PROJECTIONS OF SURFACE TEMPERATURES

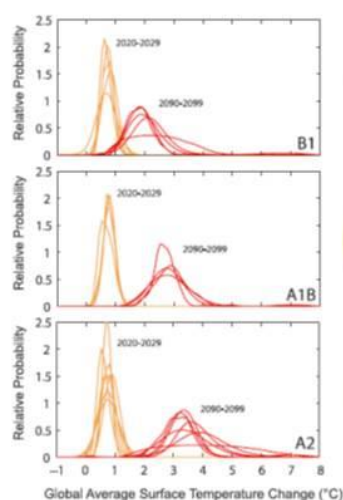


Figure SPM.6. Projected surface temperature changes for the early and late 21st century relative to the period 1980–1999. The central and right panels show the AOGCM multi-model average projections for the B1 (top), A1B (middle) and A2 (bottom) SRES scenarios averaged over probabilities of studies for the late 21st century. Therefore the distributions are narrower than those shown in the early 21st century (Figures 10.8 a

10

This risk of overconfidence is well known and well founded.

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

Not necessarily wider: they may narrow and shift under better models...

The IPCC itself might say this a bit louder/earlier: What space-time scales are realistic as a function of lead-time? (Focus on robust, but discuss inappropriate use.)

15

Leonard Smith

So this is a slide I didn't show yesterday. This is from the AR4. I think this might be the summary for policy makers, yes, and these are distributions of some temperature that is relevant to climate change. And some of them have really long tails. And on page 797 there is this nice comment about structural uncertainty, and it says that these limitations imply the distributions of future climate change from ensemble simulations are themselves subject to uncertainty and would be wider were uncertainty due to structural model errors accounted for. This basically says that every other distribution in this giant report is too narrow, unless you are interested in the models. If you are interested in the real world the distribution should be wider. You know, I am... so they said this. They did say it on page 797. But that is not quite what I would call transparent. I mean, somehow the idea... this is like hitting the wall, right? This is like the ball hitting the wall. There is a chance the ball would hit the wall put it in a footnote halfway through a two thousand page report. So somewhere, I don't know where, it doesn't really matter if they... who they cite, what matters here is somehow opening eyes to the limitations of what you are doing. And those limitations are going to be there.

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists)

Think Harder

Be as Transparent as you can be.



LSE UUEM


Royal Society

Thur PM

14 Sept 2017



Leonard Smith

I would rather have them in my paper, even if I put them in a footnote, I would rather have them in my paper than have them in a comment on paper, that Erica writes, commenting on my paper. That means **be as transparent as you can be**. If it is a theorem, say it is a theorem. If it is something about the world, say it is something about the world. Don't claim the theorem as support for something about the world and if you use it as moral support, say 'Oh, by the way, this is only a theorem'.



Laplacian Demons

$P(x | \text{data}, I)$

Laplace's Demon (1814)

- 1) Perfect Equations of Motion (PMS)
- 2) Perfect noise-free observations
- 3) Unlimited computational power

$P(x | \text{Data}, G)$ G is complete True knowledge

Demon's Apprentice (2007)

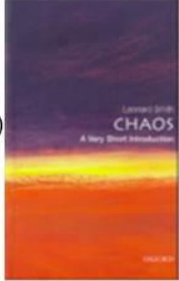
- 1) Perfect Equations of Motion (PMS)
- 2) Perfect ~~noise-free observations~~ (Noise Model)
- 3) Unlimited computational power


$P(x | \text{data}, g_t)$ g_t is imprecise but True

Demon's Novice (2012)

- 1) ~~Perfect Equations of Motion (PMS) -~~
- 2) ~~Perfect noise-free observations~~
- 3) Unlimited computational power

$P(x | \text{data}, g)$ g is useful approximations of g_t



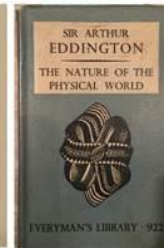


LSE UUEM Royal Society Thur PM 14 Sept 2017 Leonard Smith

La Place's Demon. Here we end up with three different probabilities in play. There is only one sort of probability we end up with, the sort of ideal probability where we have a precise notion of the exact initial condition, somehow, and the exact laws of physics. So in this case the laws govern they don't describe. I would normally say they describe, and the demon can actually give us a probability distribution we can use for anything and you can't do any better. The apprentice gives us a probability distribution we can probably use for anything, but in fact if you could get better observations it could be better. And the novice, well, the point of the novice, who actually is in the book. I didn't name him so this is really 072. I don't know what this is, right? This a probability conditioned on false and unfortunately I couldn't find my alien slide. But the next slide, it should have a picture of all those eggs, dark in the background, the music comes on, you don't need to think about something really not good is going to happen if you start talking about probability distributions conditioned on false, unless you are a mathematician, in which case you make this one of your assumptions and it is no longer false. But then again you have left 'relevant' and moved back into 'mathematics-land'.

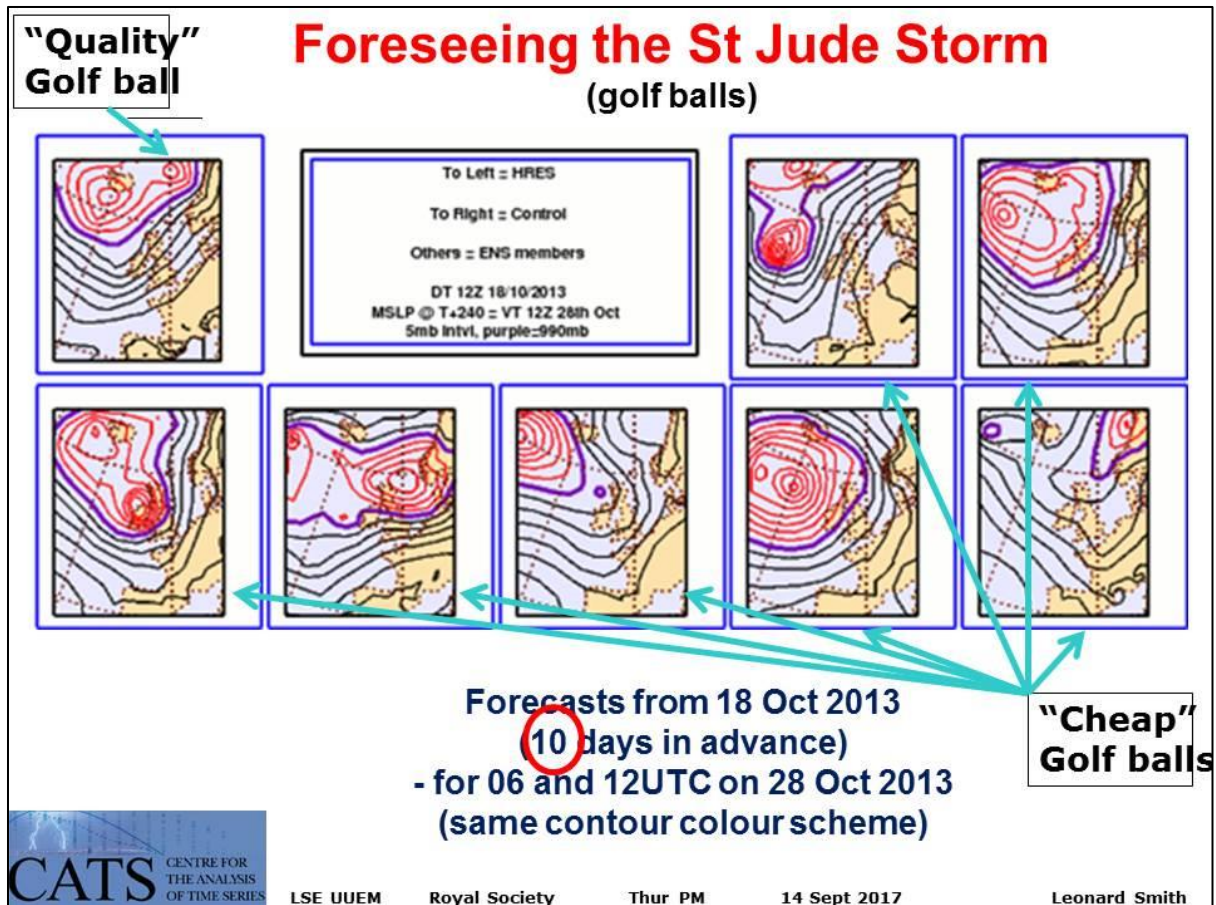
Moving on from “false ideas” is Progress!

plan, is a strong hint to alter the path.
It means that we have been aiming at a false ideal of a complete description of the world. There has not yet been time to make serious search for a new epistemology adapted to these conditions. It has become doubtful whether it will ever be possible to construct a physical world solely out of the knowable—the guiding principle in our macroscopic theories.



It appears Eddington's remarks apply even to deterministic systems!

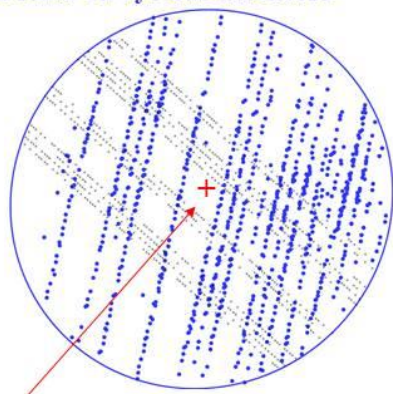
So this was Eddington's remark: "you know it has become doubtful whether it would ever be possible to construct a physical world solely out of the knowable the guiding principle of our macroscopic theories and he suggests that we were aiming for this false idea. So again, he was talking about quantum mechanics. But this structural uncertainty issue I think takes us back into a room without a clock anywhere in the room, takes us back into the same sort of issue where a lot of the things we are trying to do may not be possible. And you are never going to be able to give your thesis examiners and the referee the ability to do the things they could have done if we could say exactly what the weather was going to be. If it was possible to say exactly what the weather was going to be three weeks from today. It is always going to be less than that. Did David talk about the 2006 paper? His first Climateprediction.net paper at all? Did he talk about the reviewers? So we submitted the same day as... we waited for the Met Office. So we submitted the paper the same day and one of the reviewers said, "You can't talk about relative frequencies, you have to talk about probabilities". So Dave's paper - both papers I think said - the relative frequencies of this climate variable, because they were conditioned... it was basically assuming a model was true. It was the frequency of the model run, the next model run. We were really good at predicting this number the next model run from climateprediction.net which would come back in, and we refused to change. They wanted us to say 'probability', we insisted on 'relative frequency' and the paper Dave was lead author on was delayed by six months. But it appeared. So, there is a cost here, but this clarification somehow, it depends on what you really want to do. Do you want to do something someone is still interested in in twenty years or do you want to get a paper out in time to get your next job? That is not an easy decision, but it is one you have to think about. And one of the best ways to get both is to think of something that is new and almost as cool.



This is the St Jude's Storm which we saw briefly, but we didn't really look at. This is ten days in advance and these are golf balls. These are sort of low resolution ECMWF models running ten days in advance and look at this - there are storms! What could we actually do that would allow us some insight into the future even if we can't get probability distributions? Why can't we get probability distributions?

Implications of Structural Model Error (II)

Cartoon of
Points on system attractor



An Observation
Points on model attractor

The qualitative behaviour of trajectories (attractors and dynamic manifolds) change drastically under small perturbations to the mathematical structure of the equations.

Thus the ensemble cannot be initialised to be consistent with the natural measure of the system (if such a thing exists).

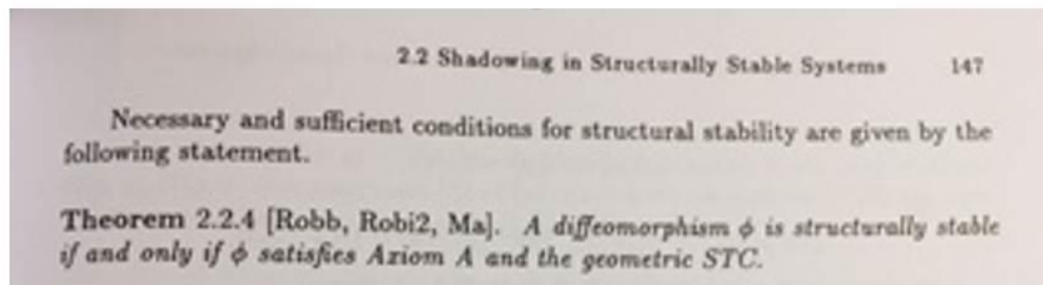
Thus even over lead times on which the model can shadow the system, one cannot construct a reliable probability forecast.

And for high dimensional systems, one cannot expect structural stability.

And “high” means greater than two.

This is a cartoon that I think I failed to show yesterday, but I will show it again even if I did. The idea in this dynamical systems theorem that I quote is just that you have this picture that we are in this high dimensional space and there are actually points distributed on the systems attractor. Those are these blue points. And the normal idea is that we want to take an observation and from that observation we put probabilities on these blue points. And then we take our ensemble and we move them forward and we see where it goes it. It is pretty simple, right? We choose the initial conditions to be consistent both with the observations which has probability of zero occurring, but also be consistent with the system in terms of this long term dynamics.

Confidence in the Extreme Weather of a Climate



The lack of structural stability implies we cannot identify states of a (high dimensional) system that are both consistent with the obs and with the long term dynamics (“on the attractor”)

This does not place any finite limit on how long the model can shadow.

And our current models sometimes do storms rather well...



13 Sept 2017 Predictability Probability and JEDI Insight University of Reading Leonard Smith

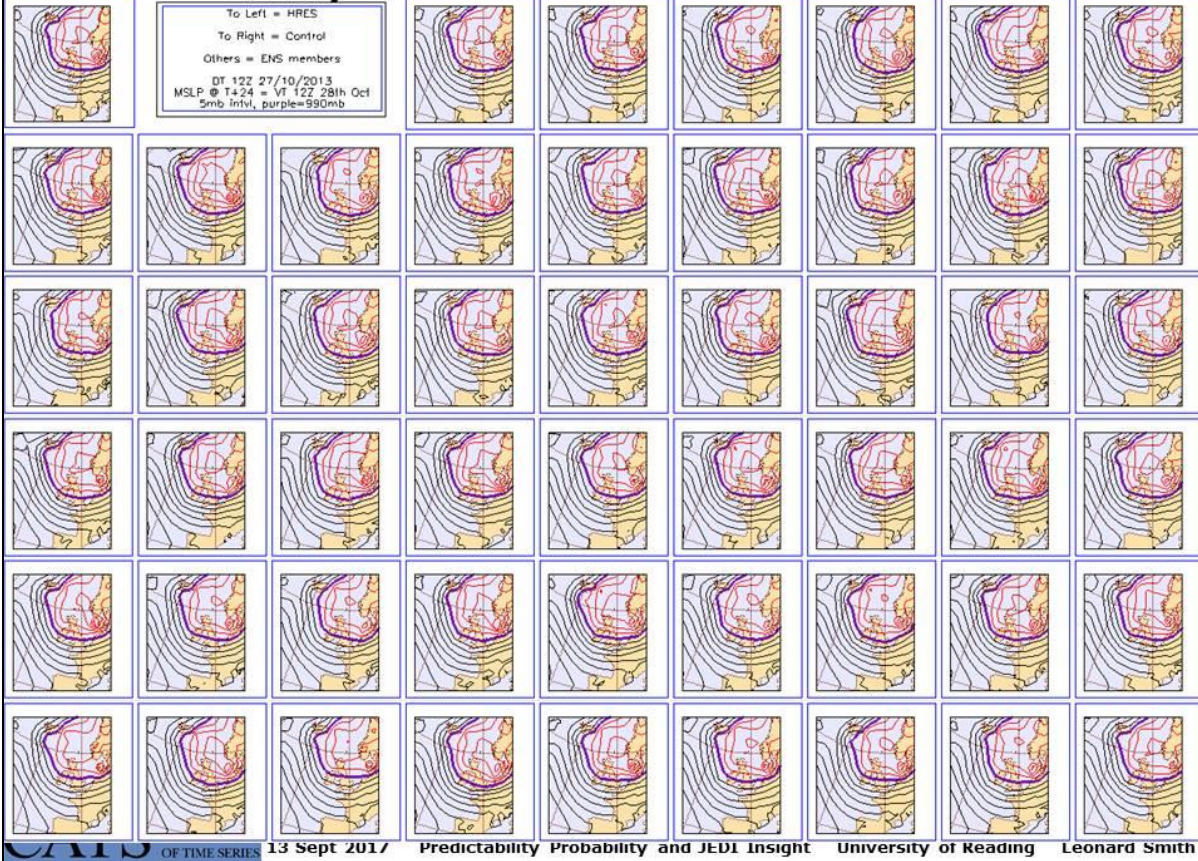
The problem is these little grey dots are actually points on the model’s attractor. I don’t even know if the system has an attractor. It only runs once, right, something hits to make the moon, so before that everything changes, then afterwards the sun vaporises the planet and we don’t exist anymore. The sort of limits we have to take to define the mathematical objects like an attractor require we take limits to plus infinity and minus infinity, and that doesn’t make sense any more. So the best we can do is to take the observations and put a distribution on these light grey dots. That would be the wrong probability distribution to start with, which means we are going to have the wrong probability distribution as we go forward in time. Let’s try to do something else. Well this theorem says that we can’t... what I just said, but it doesn’t place a finite time on how long the model can shadow. And sometimes our models do really well.

This is the twelve hour forecast. This is twelve hours before the storm and there are storms everywhere, and they pretty much look the same, and that is the high res run. So this is the expensive golf balls. These are the cheap golf balls and they all look pretty good.

51 ensemble members forecasting t=0, each 12 hours ahead



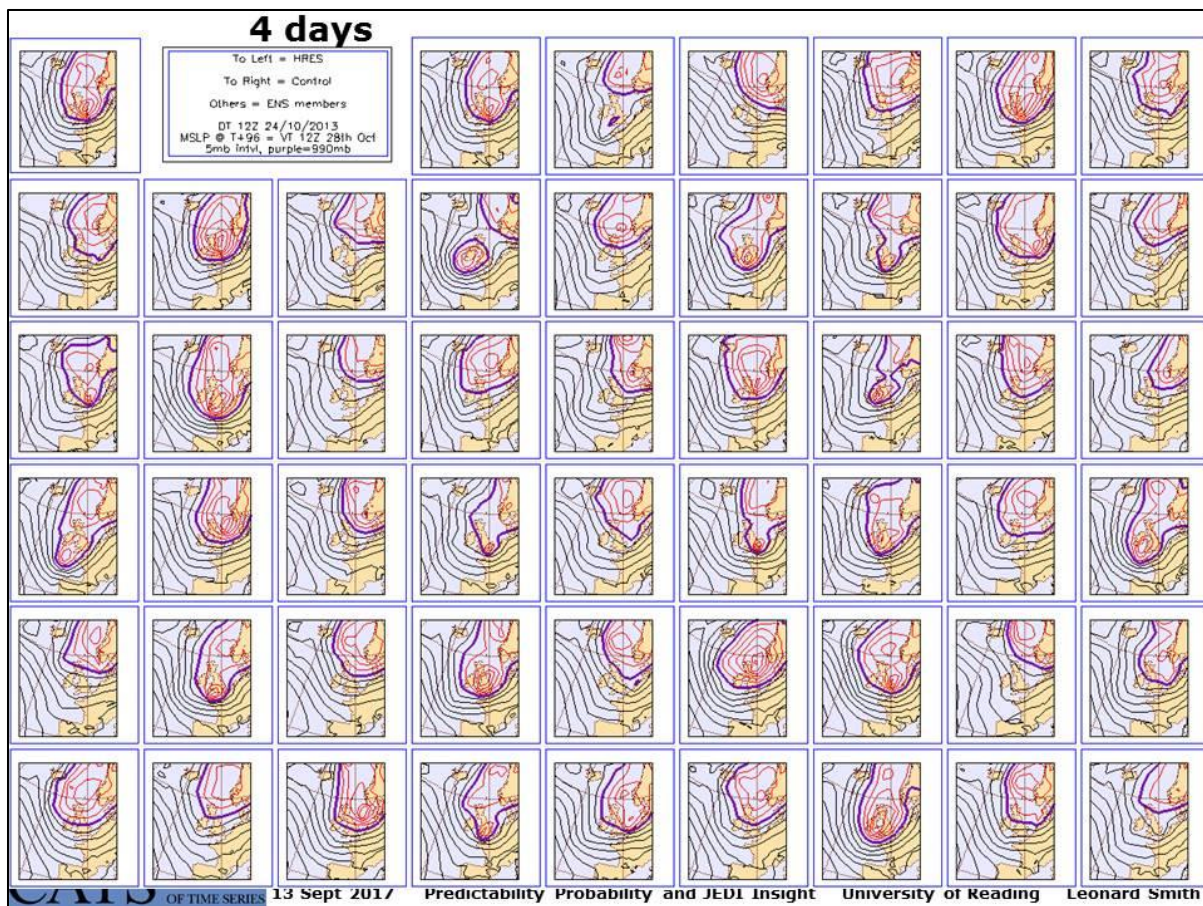
1 day



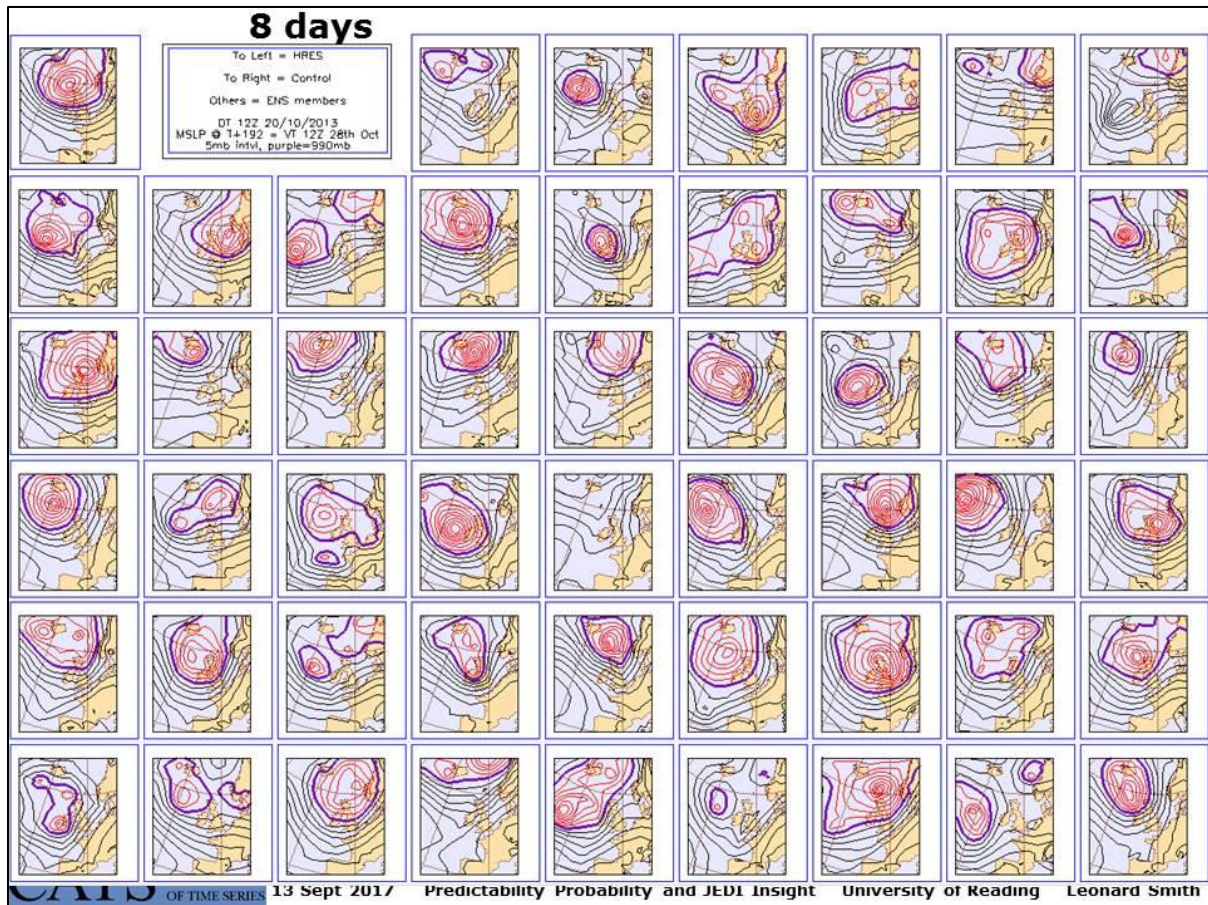
And this is a day ahead, and lots of storms.



And this is two days ahead.



This is four days ahead. Now people start talking about an error doubling time of two days in the atmosphere.



So by four days the errors are pretty big and by eight days they are *really* big, and people start saying you can't forecast much longer than eight or ten days, maybe two weeks. And there are still storms!

And this is fifteen days - and there are still storms!



Actually, this one is a pretty amazing storm. It is a big storm and it is in the right place. So fifteen days ahead we actually realised there was the chance of a model storm, maybe a real storm. The expensive golf balls stop before this and I would have gone on further but on Day Sixteen they get even cheaper golf balls, so there is a lack of continuity. So what can we do? So this is really good and I checked, there are not always storms at Day Fifteen. So basic statistics. You still have to do the basic statistics, right? But we couldn't evaluate this as the probability of a giant storm. I mean, if we just counted the number of storms and then divided by 51, this is a bad answer.

Early Warning of Events with GLIMPSE



Spot an interesting ensemble member at large lead times.

Then sculpt an ensemble (sampling over everything) to “enhance” the target.

Then monitor the implied physical development:

Does a plausible model-trajectory lead to the phenomena of interest?

What observations (now) are most informative regarding that plausible (model)event?



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

13 Sept 2017 Predictability Probability and JEDI Insight University of Reading Leonard Smith

So let us do something else. Let us look at an interesting ensemble member and let us say that something that happens in the future that looks really cool or really bad. Whatever turns you on. It could be a nice thing, it can be a horrible thing. You decide what you want to look at. And then let us sculpt an ensemble initial conditions to see if we can actually make that storm bigger, or that beautiful sunny day with good surfing better. And then let us monitor the implied physical development, let us see what the model does. The model trajectory that leads to that phenomenon, what does it look like? And then let us see: does that happen? Can we make observations to see if this is really happening? And if it is a bad enough event can we actually take action now? Or if it is happening in Day Twenty One maybe we just wait. And is it still happening at Day Ten? We are starting to think about what to do if that thing was to happen. So the National Grid, for example, whenever it is really, really cold a transformer, maybe two, blow up and these things are as big as this room and they are really expensive. And there are maybe fifteen really old ones left in the country and they don't know which ones are going to blow up. So what can they do? Well five days ahead of time they can have the guys in the warehouse move the transformers from the back of the warehouse to the front of the warehouse. This costs almost nothing because those guys are working in the warehouse anyway. And yet if a transformer blows up they save a lot of time in getting it out.

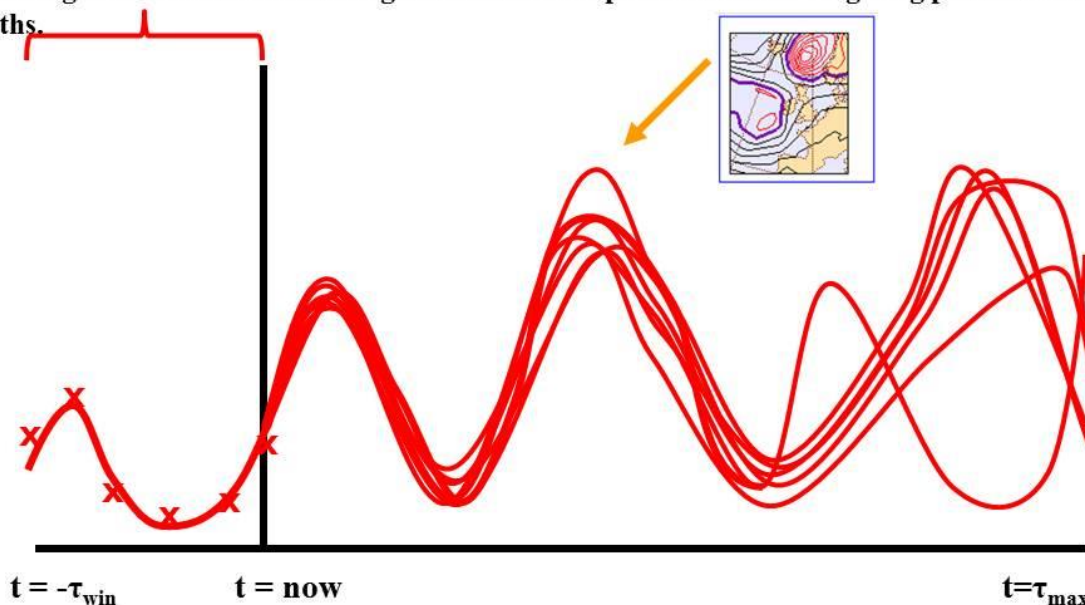
When I was talking to my brother... my brother who was just in Irma, they have trucks from all over the southeast in Jacksonville now. He had Tennessee Valley Electric Authority trucks in his neighbourhood. Those trucks didn't drive from Tennessee after Irma, they came down before the hurricane went through Florida. So they prepositioned things. That was much more expensive. And my best story is about filling the refrigerator on an aircraft carrier, but that takes too long. But this is something that is pretty cheap that you can do a lot of times. You don't need to know the probability.

Gauging Limitations of Imperfect Model Prediction with Sculpted Ensembles

Large Ensembles in a 10^7 dim space are infeasible.

Linearized dynamics in week two are irrelevant.

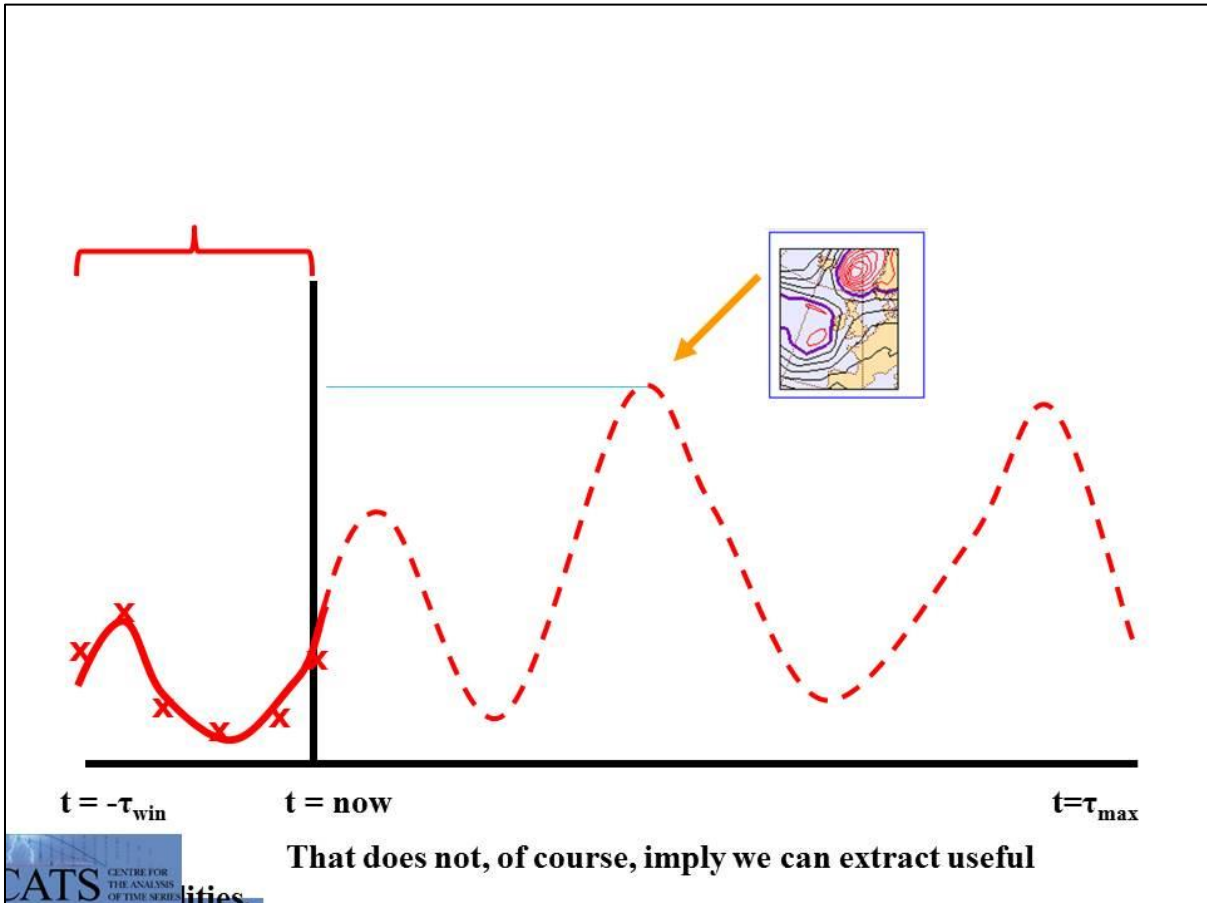
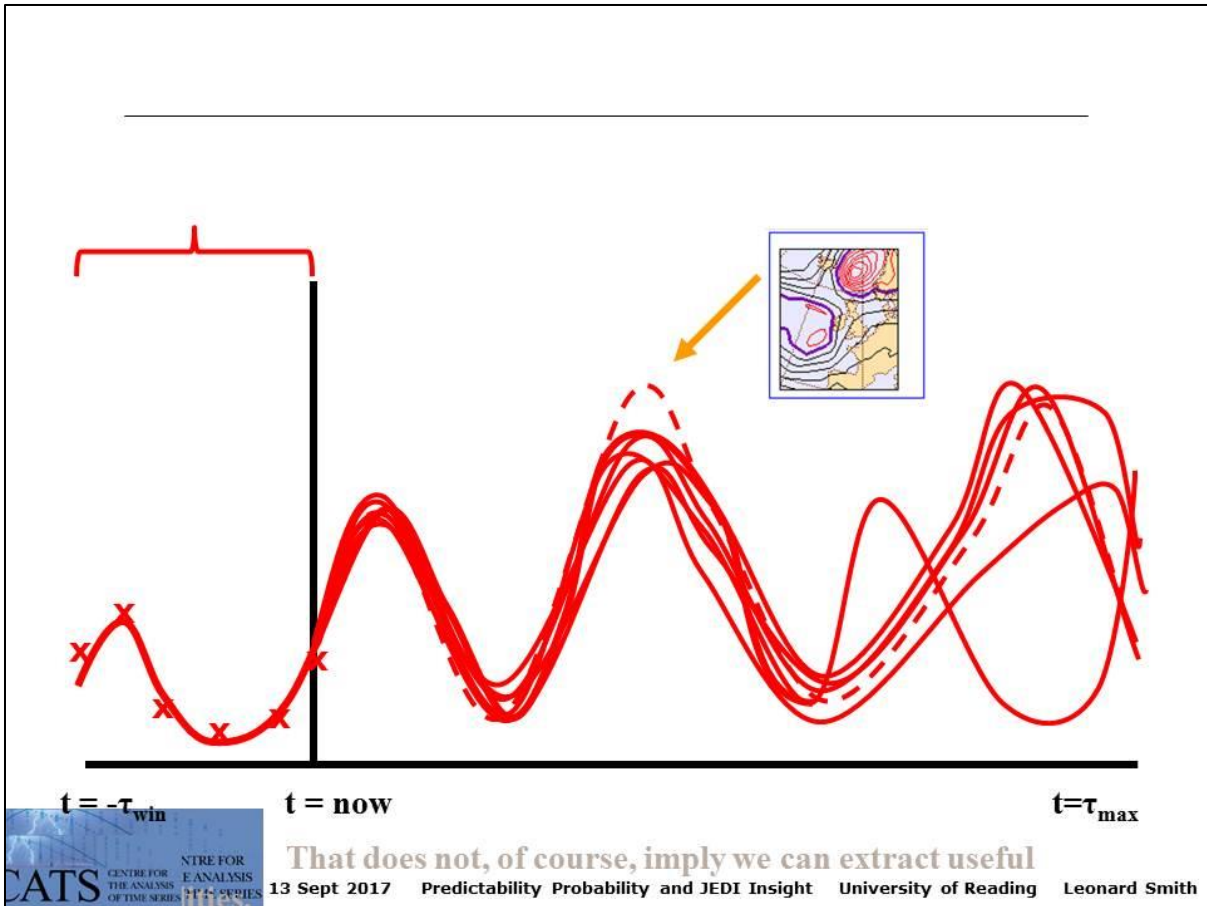
One might GLIMPSE interesting events with sculpted ensembles targeting plausible model paths.

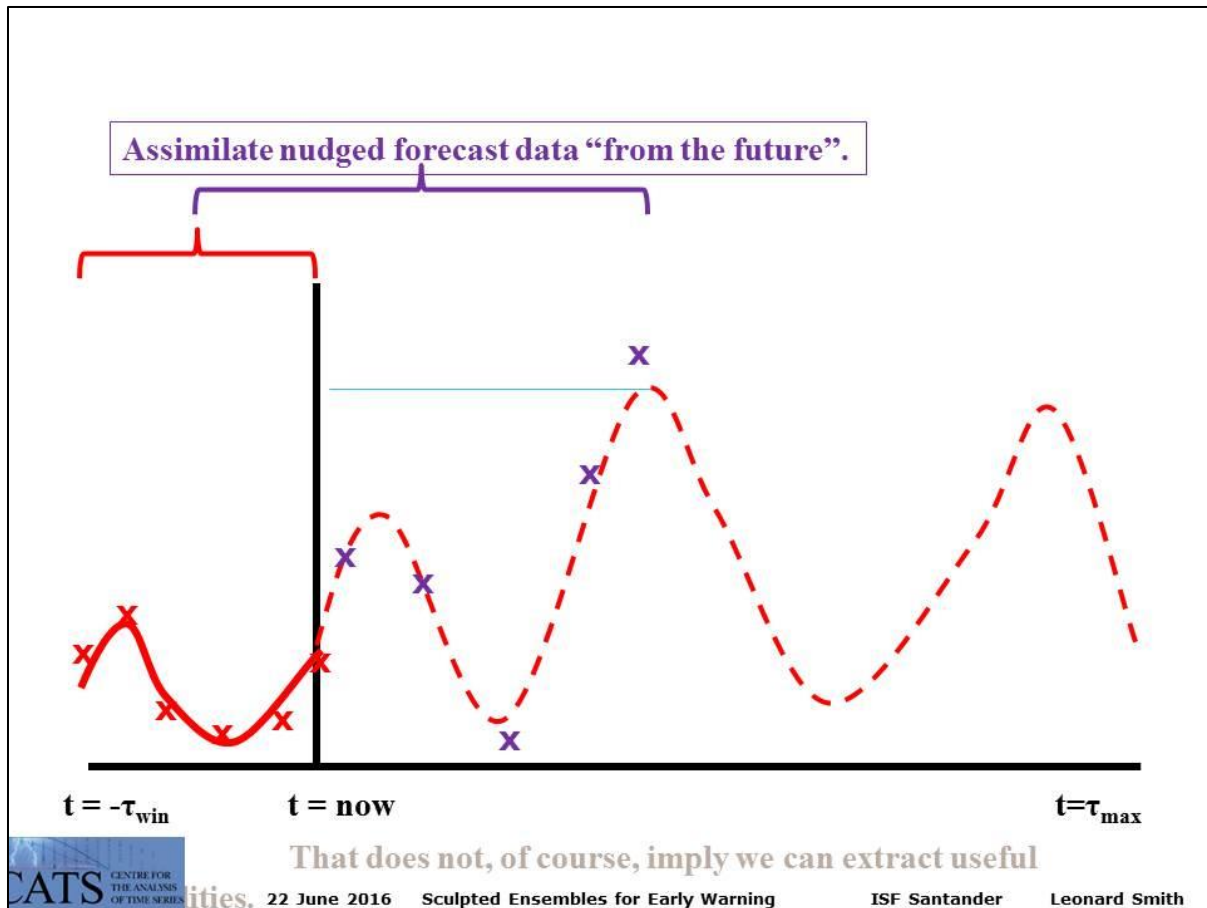


CATS
CENTRE FOR
THE ANALYSIS
OF TIME SERIES
ities.

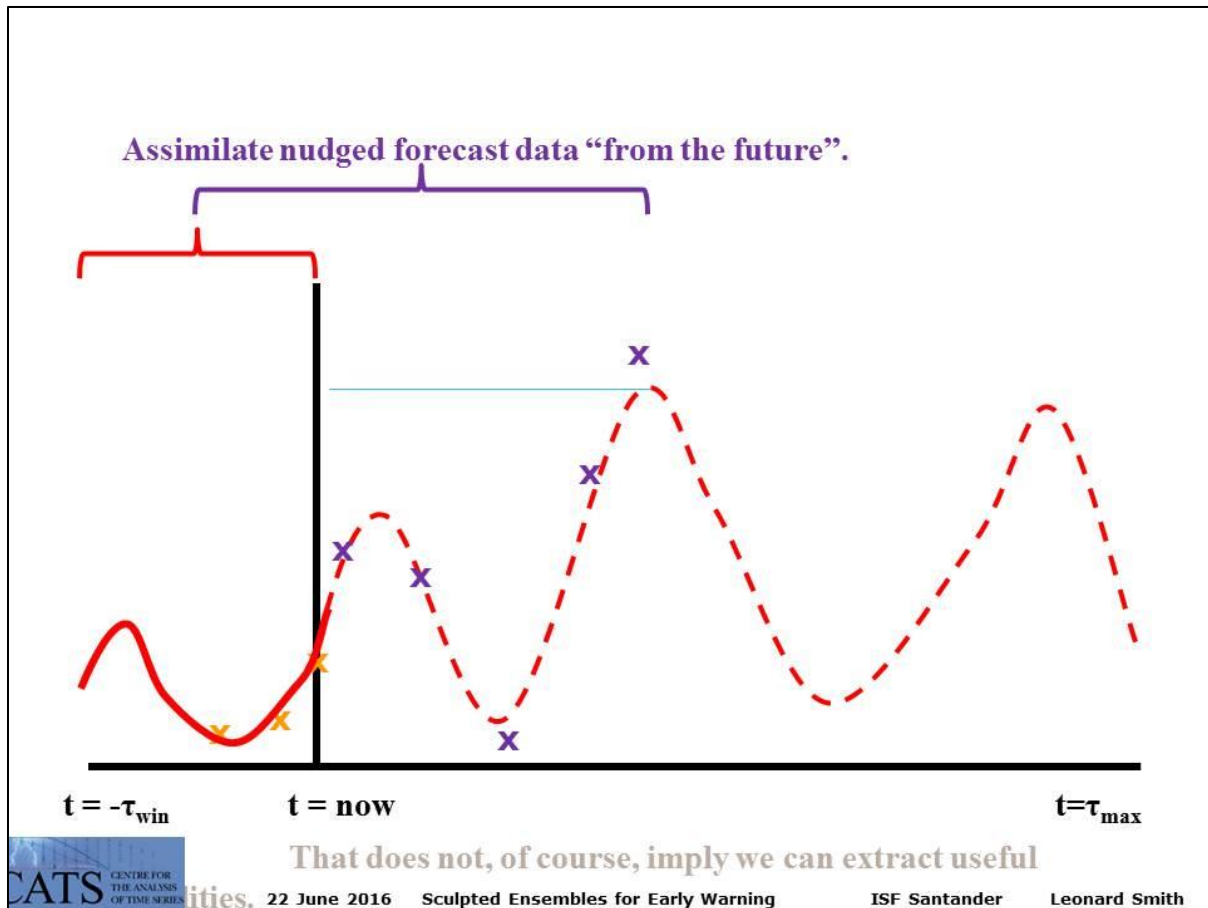
That does not, of course, imply we can extract useful

How does it work? (This example is from Ed, I think.) This doesn't have to be a forecast. You are assimilating data here, we launch this ensemble and hey look! There is something big. So we identify that storm or that good event. So that is sort of cool. So this could be, you are just working on a parameterisation scheme and one particular parameterisation does something interesting that you are worried about, and you can study that. But in the time series case we focus on that particular ensemble.

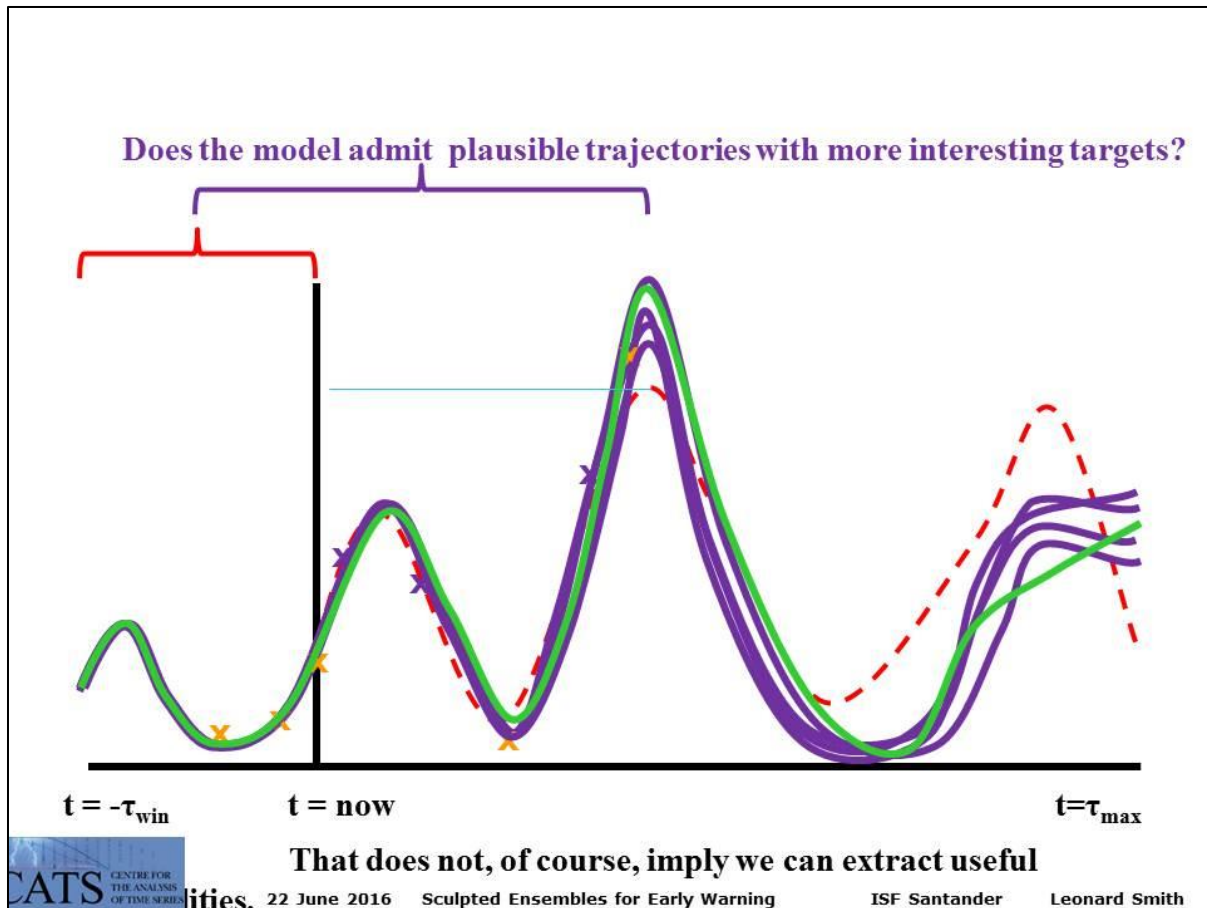




This is now. We make some fake observations out here in the future, and we have a really cool data assimilation scheme that allows us to assimilate the future. That is with Du, and Kevin Judd and others. We then take those observations and we actually look, we still want to be consistent with the past, but we actually assimilate data all the way up to here. Maybe we nudge the storm a little bit deeper, or the waves a little bit taller, the sun a little bit brighter.

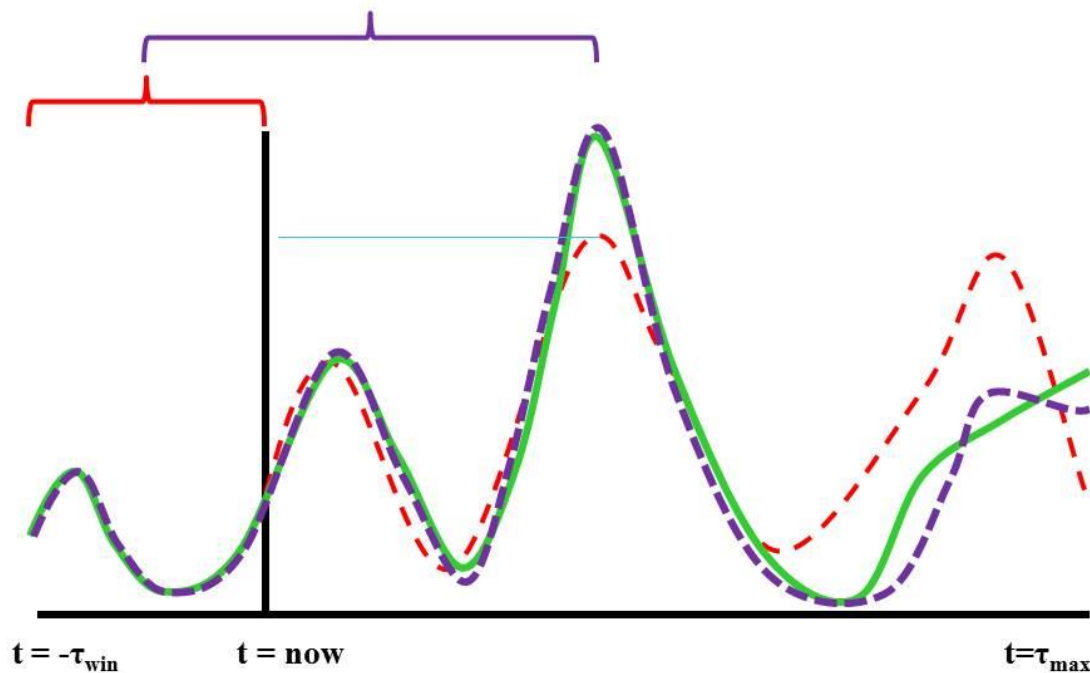


And we find another...



We ask the question: does the model actually admit a trajectory that does even more of what we want or what we fear? And then we look at that trajectory. So what we have done is, having seen this guy in our forecast, we say “Hey, what is the worst that can happen?,” well not the worst but, Can we make this worse? Or better? Can I make it sunnier on my wedding day? Can I make the waves bigger on my surfing weekend?

Early Warning of Events with GLIMPSE



Early Warning of Events with GLIMPSE

GLIMPSE

Spot an interesting ensemble member at large lead times.
Then sculpt an ensemble (sampling over everything) to “enhance” the target.
Then monitor the implied physical development:
Does a plausible model-trajectory lead to the phenomena of interest?
What observations (now) are most informative regarding that plausible (model)event?

GLIMPSE aims to provide Just Enough Decisive Information to allow:

General Early Alerts

Cost-effective protection: Insurance on warning

Cost effective Regulation : Turn off plant rather than “survive operating”

Low cost Rerouting: Logistics

and so on.

GLIMPSE does not overcome model inadequacy (but can help identify it!)

This is a viable forward path for simulation based forecasting which allows us to write challenging code!

And so the aim again is to provide Just Enough Decisive Information to allow early alerts for those people who could use them. You don't want to scare people too much. You need to think a lot about how you release the information.

Cost effective protection. You could have some sort of insure on warning so that when this alarm goes off the insurance company pays out so that you can afford to buy the plywood to put on your windows. If you have been living in Florida a while you already have the plywood in your garage with the holes drilled and everything. Anyone who has been in Florida for a generation or two, there is no rush to buy plywood in the days before a hurricane. It is already there. So it doesn't cost much. It is easy to do. It doesn't stop a tree from falling on your car.

Cost effective regulation. Well for nuclear power plants the cost of following today's regulation for climate change would be incredible, but you can turn off a nuclear power plant in two and a half days. And probably you can turn off even faster if that was important. So rather than have the plant regulated so that the storm comes out of nowhere like an earthquake you could actually have it regulated in the way that you turn it off. And that saves enough money to build four or five in different places.

Low cost rerouting for logistics. There are just ways of viewing this information that would be really fun. But more important than that we can do this. We can do this today. We can't overcome model inadequacy, we can't have probabilities, but we don't have probabilities today. And when I give this talk to a meteorological audience, or when Du gives this talk or maybe even Ed, they say "but you don't have probabilities. We are not going to use this if you can't give probabilities," but we don't... this is another very common claim. The fear of losing something we don't have. The fear of losing something you can only get under what Eddington would have called this "wrong approach" and it still allows you to write cool code.

Focus on YOUR question!

The **tools** are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists)

Think Harder

Be as Transparent as you can be.

Find cool questions (useful and solvable) and serious people.



13 Sept 2017 Predictability Probability and JEDI Insight University of Reading Leonard Smith

Again, find cool questions that are useful and solvable and work with serious people. So work with serious people that are older than you are. Find really good colleagues to work with like these guys, who make half my slides. And Ed was making one earlier in this talk. Really this is just a nice way of doing good science with mathematical models. And you can still write cool code.

Simulation Modelling is a Driving Force in Science.



It need not fall to the dark side.



13 Sept 2017 Predictability Probability and JEDI Insight University of Reading



I don't know if I showed this one. Did I show this one here yesterday? This is another one of these Darth Vader as your thesis adviser.

"Luke code with me,"
"Why?"
"Because it is the only way,"
"Why?"
"Because,"
"Because why?"

There is a sense in which he is right. The tools are really important and the power that doing mathematical modelling in physical science gives us is huge, unless we try to do something impossible and then eventually learn to do something impossible is to go over to the dark side.

Early Warning of Events with GLIMPSE



Spot an interesting ensemble member at large lead times.
Then sculpt an ensemble (sampling over everything) to “enhance” the target.
Then monitor the implied physical development:
Does a plausible model-trajectory lead to the phenomena of interest?
What observations (now) are most informative regarding that plausible (model)event?

GLIMPSE aims to provide Just Enough Decisive Information to allow:

General Early Alerts

Cost-effective protection: Insurance on warning

Cost effective Regulation : Turn off plant rather than “survive operating”

Low cost Rerouting: Logistics

and so on.

GLIMPSE does not overcome model inadequacy (but can help identify it!)

This is a viable forward path for simulation based forecasting which allows us to write challenging code!



13 Sept 2017 Predictability Probability and JEDI Insight University of Reading Leonard Smith

But you don't have to do that if you can find something that is possible, useful and solvable, or just solvable, then you still get to code. And you even get to code more interesting stuff.

I want to say a couple more semi-factual things. There is lot of talk about probability, both in science and maths and in regulation. And this is sort of there are four-ish different kinds of probability – these are all due to I. J. Good. One is this topological probability. This is sort of like flipping a fair coin. And this is either true or false. He used something like the thousandth digit of pi, the probability that the thousandth or the millionth digit of pi is even is not fifty-fifty. It is either zero or one. We just don't know the answer. But it is a well-defined question. As soon as you say “a fair coin” you are done. You could say “spinning a fair coin” as long as you knew what that meant.

Four-ish Kinds of Probability (IJ Good)

$$P(x | \mathbf{I})$$

Rational Decisions I. J. Good (1952) Journal of the Royal Stat Soc. Series B (Methodological) Vol. 14, No. 1, pp. 107-114.
Good Thinking I.J. Good (1983) Dover.

(o) **Tautological Probability.** A probability $P(x|H)$ the value of which is specified in the definition of H . (“a fair coin”, H is “a simple statistical hypothesis”)
Arguably $P(x)$ is conditioned on nothing beyond the problem statement.

(i) **Physical Probability:** $P(x | \mathbf{I}_{full})$ “True probability” of x .
(The probability held by Laplace’s Demon/Infinite Rational Org)

(ii) **Subjective₁₉₅₂ Probability:** $P(x | \mathcal{G} < \mathbf{I}_{full})$ probability of x given information \mathcal{G} which is true but incomplete. \mathcal{G} is a subset of \mathbf{I}_{full}
(The probability held by the Demon’s Apprentice/?semi-finite Rational Org?)

(iii) **Subjective_{My} Probability:** $P(x | \mathbf{I}_{my})$ my attempt to estimate the subjective₁₉₅₂ probability, given imperfect models, inexact observations, finite computational power and tidbits. **Key point: I can know my $P(x | \mathbf{I}_{model})$ is not mature.**

A **Mature probability** is not expected to change without some new insight or additional empirical observation (even given vast increases in computational power).



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

Physical probability - well this is the true probability, and as Good says it may not exist but it is still sort of useful to talk about it as if it did exist, because this is the thing we could use, and knowing we don't have it yet is a useful thing to know, even if we don't have it yet because it doesn't exist. When he was talking about a subjective probability he was talking about the Demon's Apprentice. He was talking about an attempt to estimate *that* and today this personal subjective probability is just... What does Ed think? I mean I like Ed, but who *cares* what Ed thinks? Why should... it depends on what it is. He has been barred from three casinos. I have yet to be even escorted from one of them. So maybe he actually knows something, but if it doesn't have to do with casinos, if it doesn't have to do with something he knows about... And if it has something to do with something he knows about then he is probably trying to do *this* because he wants to get to *that*. If it is a field in which our models are misinformative then a modelling expert may not know much more than stupid things to do. So rather than look at these, I would suggest we look at what I would call a mature probability. This is just one we don't expect to change without some extremely new insight or new cool observation. So even if we had super faster computers we expect to be able to go with this kind of probability. If you ask domain ...people who are doing simulations, if you are working on a subroutine that has only three lines of code to say what termites are going to do in a different climate one equation one equation, one 'if' statement, right, then probably it is not a mature probability and you can learn that really quickly and the way it is treated in decision making is different. And it still may be the best we have.



Convergence in Distribution



Obtaining target trajectories can be difficult, and identifying shadows expensive...

An alternative approach is to ask if our in-hand models have converged in distribution, even as their individual trajectories differ greatly.



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

Someone must have done this already, who?

So this is Oberkampf again. And how do we go about figuring out if the probabilities are mature?

Mature Probabilities $P(x|I)$

A mature probability is not expected to improve without additional observation or new theoretical insight. (An nontrivial change in I)

If the fidelity of a simulation model is constrained by technology (as when you know exactly what you would do with more compute power, and it is NOT to run massive ensembles/**emulators**), then probability distributions based on simulations from that model (or family of similar models) are not expected to be mature.

Rational action is constrained only by mature probabilities (or those believed to be mature).



(Generalised from IJ Good's "Dynamic Probability", as when output from a chess program must be used before the algorithm completes.)
LSE UUEM Royal Society Thur PM 14 Sept 2017 Leonard Smith

I mean you can sort of think they are, but you can know they are not. And then just let me just talk about this graph for a minute.



Convergence in Distribution



Obtaining target trajectories can be difficult, and identifying shadows expensive...

An alternative approach is to ask if our in-hand models have converged in distribution, even as their individual trajectories differ greatly.

The **Discrimination Information** (Kullback's term for RE) for three model

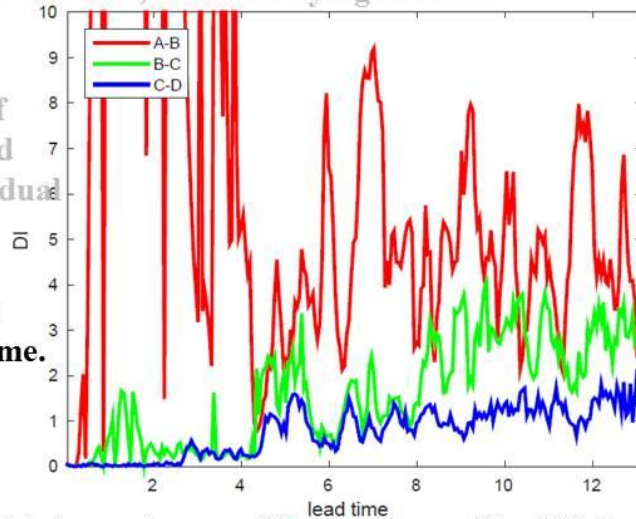
$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx,$$

where p and q denote the densities of P and Q .

The y axis is in **bits!**

Going from **model A to model B** yields huge changes (?improvements?) ~10 bits. From **model B to model C** significant changes (~2 bits even at short lead times)

While **model C to model D** may have converged for $t < 2$?
Someone must have done this already, who?



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

The point of this graph is really to say I have a model and I build a new model and I compare the two. Or there is a parameterisation and someone runs a new model, hopefully changing only my parameterisation. So probably they change a lot of stuff, but maybe if you are working on this subpart the actual way that you move forward with a model isn't to run the whole thing with everything. You actually do go through and you change things individually and you look for the sensitivity. Wouldn't you learn a lot doing that? You would learn a lot about the model. You get to write a bigger grant because you have to run the model more often, and you generate more CO² but you actually might move forward in a way if you thought about how to do this. That you actually learned about the sensitivity to the different pieces. You could even leave some switches in there for jumping between them if you couldn't tell which one was better in terms of overall performance. And so that is what this is trying to do.

Discrimination information. If I run Model B, having had Model A, then what I see is that things change hugely. The probability distribution changes by a factor of two to the ten a thousand times. So I am looking at different times as I move forward in time. Going from Model A to Model B, I know Model A is not mature because by running a better Model B everything changes. Because of course Model A is usually cheaper so I have a really good pdf of Model A, not such a good pdf of Model B. I build Model C and I compare Model B to Model C. And look, it is still factors of eight and ten and thirty two. So this is telling me that Model B, well you know, it is not mature either. So I run large ensembles under Model C and I compare them to Model D. And I get this tiny little value out to about three days. Well that is sort of saying that Model C and Model D are pretty much consistent down here and then they drift away. But I am starting to get to the point where, at least for this particular place, my model might actually be mature. I mean, I would probably want to do some more. But here I know it is not. Here I have the shot of saying "Hey, whatever I am improving here I

have sort of got it". If I am looking at things in this range then I probably wouldn't have integrated out this far. I am still waiting for Model E. Has it run yet?

[Possibly.]

I am still waiting for the results from Model E, but what I would hope is that slowly I get this consistency going out farther and farther until it is sort of reaches the lead time that I want. And again I can't know that well there could be an unknown unknown or a mutually known neglected between these models, but at least as far as I can tell it is mature. I don't know before I go to talk to someone that I don't believe this model. And I have actually had more success talking both for neutral policy and for profitable consulting. Talking about the limitations is really a good thing. I am surprised more people don't do it (or maybe they do).

Mature Probabilities $P(x|I)$

A mature probability is not expected to change without additional observation or new theoretical insight. (An nontrivial change in **I**)

If the fidelity of a simulation model is constrained by technology (as when you know exactly what you would do with more compute power, and it is NOT to run massive ensembles/emulators), then probability distributions based on simulations from that model (or family of similar models) are not expected to be mature.

Rational action is constrained only by mature probabilities (or those believed to be mature).

?Can one use distributions not (believed) mature in decision support?

**By providing the probability that the distribution in hand is misleading:
 $P(\text{Big Surprise})$**



LSE UUEM

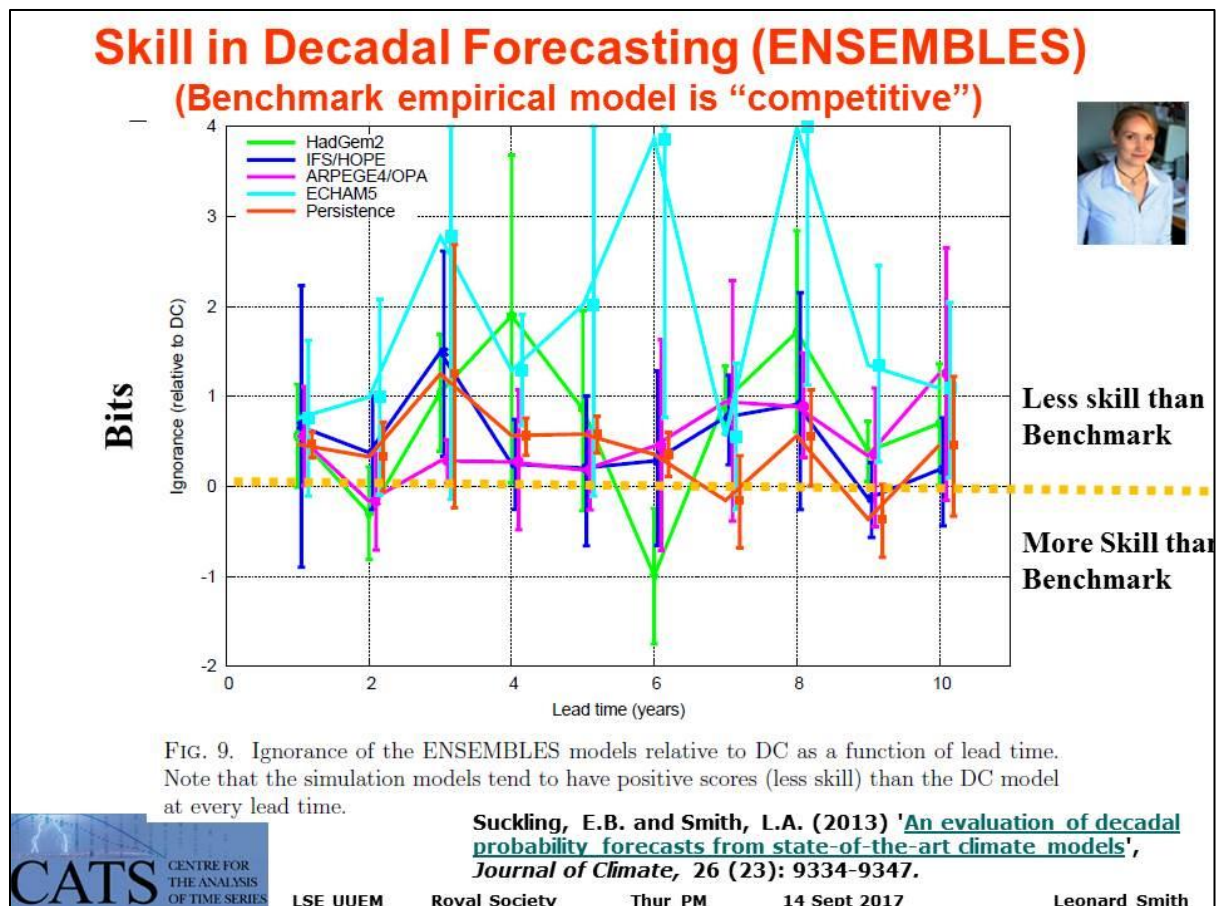
Royal Society

Thur PM

14 Sept 2017

Leonard Smith

And how do we get to this? Well even after we have gone through this thing, I mean if we don't believe the distribution I don't know how to really use them in decision support. But one way to do it would be to provide this Probability of a Big Surprise. You will be the expert. What do *you* think our inability to have a good termite subroutine or a penguin subroutine...when is that really going to kick in in climate? Or translate that to your problem? When does it matter in terms of the energy system of the country that we don't know how we are going to deal with embedded generation? That it is not going to be that everyone turns on their lights in Birmingham, turns off their lights when the clouds go away, but also that half the equivalent energy of half the generators plants suddenly appears because all the solar cells start working. We don't know. How big could it be? How bad could sea level rise be in the next seventy years?



This is a result with Emma. The other thing you could do is you could take a really expensive model and replace it with a card trick. So let us suppose I am running an extraordinarily expensive set of models to predict the changes in global mean temperature. And then (I won't ask Ed), but I will ask you to pick a card. What does it say? Don't show it! I only asked you what it said. This says "0.9 degrees". Now I take today's global mean temperature and I add a 0.9 degrees to get next year's global mean temperature. Or I could run a giant computer model, and then I subtract 0.02 degrees to get the second year, and the third year, and the fourth year maybe I do it with replacement. However, what Emma did... we built a model which just took all the paths changes in global mean temperature, selected them at random and then added them to this year and she beats in this case a bunch of state-of-the-art models that were state-of-the-art ten years ago, eight years ago. Deck of cards. I mean, it is a little more complicated than that. What she actually did would have required five decks of cards or ten decks of cards. But still, ten decks of cards custom printed (thanks to Jill) is cheaper than the CPU power that went into this ensembles project. So, forget this particular example, your component or your model is trying to simulate something alright? You are trying to understand something. Think of a really silly way to do that and then try and compare how well your silly deck of cards compares to the expensive models. So this is... in bits. So zero means that the deck of cards puts the same probability on the outcome as the expensive climate model, seasonal model. One means the deck of cards, well minus one... these are playing against the deck of cards so positive values means the deck of cards puts more probability than the climate model. In other words, let us take a value that isn't too bad, you notice these bars, these bars are resampling bars, these are the uncertainty. If the values are positive the deck of cards wins. In this case the green model...in this case forecasting six years into the future the Hadley Centre model put more probability on the outcome, but only on this case and the error bar still had zero. But is it minus one? So in this case the Hadley Centre model, on average, put twice as much - two to the one - this would

be four times as much, four times less, eight times less, sixteen times less. So if you can find some quantitative way of comparing surrogate forecast system, some simple forecast system, with a complicated forecast system, or a parameterisation or the model, you don't even have to do it against real data. If you can actually get your model to run you can try your complicated parameterisation and your simple parameterisation and run them whole model. If you can't run them whole model it gets a lot harder. Does that make sense?

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists)

Think Harder

Be as Transparent as you can be.

Find cool questions (useful and solvable) and serious people.

Have fun!

And the thing is don't be afraid of failing these tests. One of my most cited papers, two or three of my most cited papers are basically pointing out that one wants to have tests which models tend to fail robustly.

Decks of cards are fun. I accidentally took the wrong deck of cards to North Carolina during the eclipse. And I ended up with a deck of cards which was blank. I still used them. I just had a very senior guy pick the card – it was blank! It still worked. So just find a way to have fun. And having fun with serious people is easier. So I am going to walk through this very quickly and then stop.

Parameter Indeterminacy

“Selection” or “Declaration” not “Estimation”



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

The other issue that came up is this issue of things like parameter selection. So parameters.. if the model is perfect you are trying to estimate what the true value of the parameter is. But if the model isn't perfect there is no true value of the parameter. Even if the little piece of your parameterisation has an analytic result which is relevant, well that may not be the best parameterisation to put in.

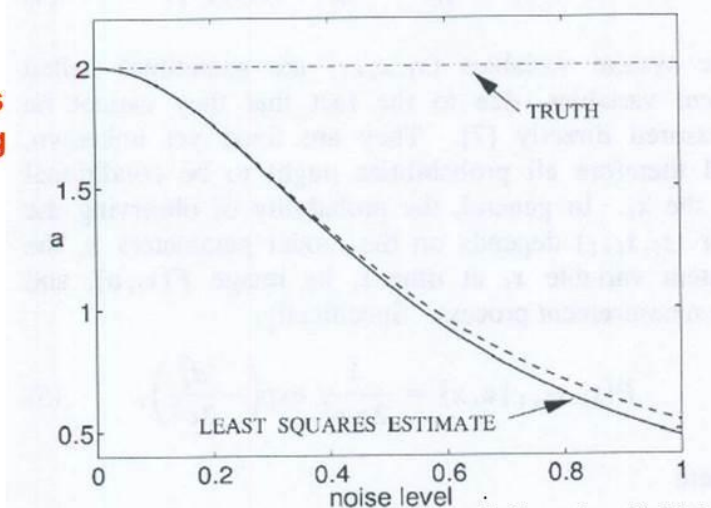
Parameter Selection in PMS

Even given a perfect model, and an IID Gaussian observational noise, parameter estimation in a nonlinear model is nontrivial.

How would you approach this task?

a) Minimize the RMS forecast error as a function of α

RMS Error is a misleading measure of skill!



McSharry, P. and Smith, L.A. (1999) 'Better nonlinear models from noisy data: attractors with maximum likelihood', *Phys. Rev. Lett.*, 83 (21): 4285

And all I want to say here is that when the model is perfect then there are many ways of finding the true parameter value. Ed can do it. Erica has done it. But... and there are cool things you can play with. And if you choose different skill scores, as long as they are proper, which Ed talked about, then they are all going to agree. But if the structure of your model isn't perfect then they will disagree. If the structure of your model is just a little bit off then there is not a true value any more and it is not estimation. It just becomes selection hopefully, right?

Parameter Selection in PMS

Even given a perfect model, and an IID Gaussian observational noise, parameter estimation in a nonlinear model is nontrivial.

How would you approach this task?

a) Minimize the RMS forecast error as a function of α

b) Look for the most similar natural measure given the noise model.

McSharry, P. and Smith, L.A. (1999) '[Better nonlinear models from noisy data: attractors with maximum likelihood](#)', *Phys. Rev. Lett.*, 83 (21): 4285-4288.

c) Look for the longest shadows.

Smith, L.A., Cuéllar, M.C., Du, H. and Judd, K. (2010) '[Exploiting dynamical coherence: A geometric approach to parameter estimation in nonlinear models](#)', *Physics Letters A*, 374, 2618-2623.

d) Look for the **best probability forecasts**?

Du, H. and Smith, L.A. (2012) '[Parameter estimation through ignorance](#)', *Physical Review E* 86, 016213.

Indeterminate parameters in imperfect models

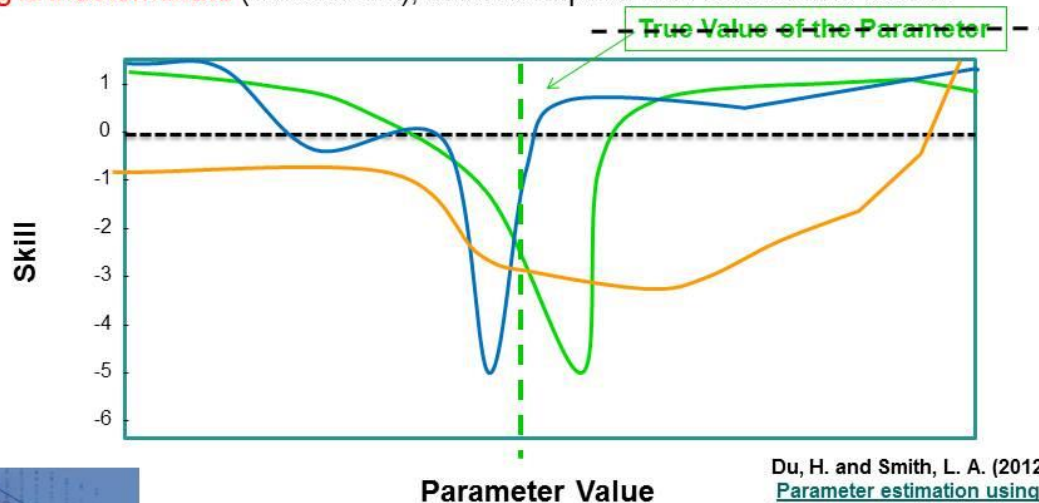


Given a Perfect Model Structure:

All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists).

Given an Imperfect Model Structure

The particular Score matters, the forecast lead-time matters, **the value you are targeting is indeterminate** (none exists), and the implied IGN reveals info deficit.

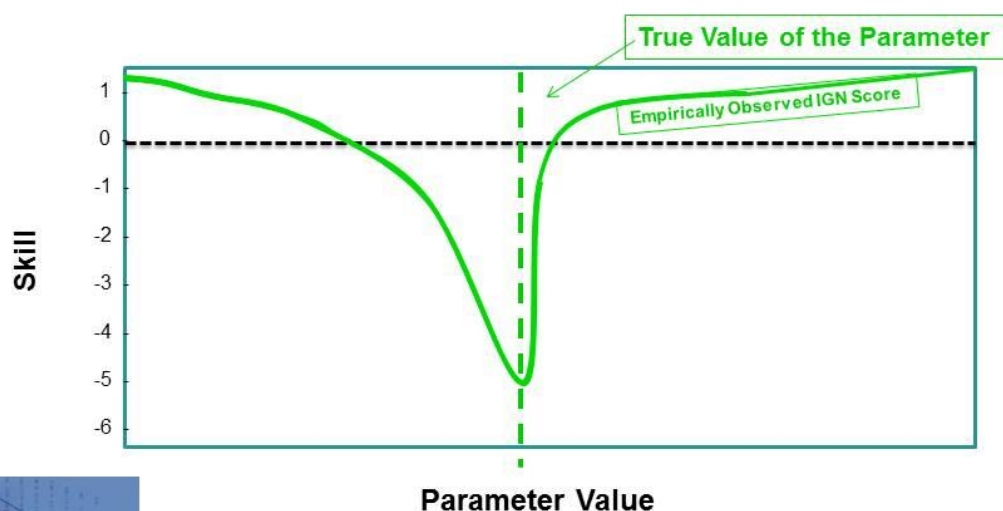


Du, H. and Smith, L. A. (2012) '[Parameter estimation using ignorance](#)' *Physical Review E*

Identifying Parameters which Maximize Information (Minimize "IGN")

Given a Perfect Model Structure: Make probability forecasts for different parameter values. All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists).

IGN = $-\log_2(P(x_{\text{outcome}}))$ Proposed by I J Good in 1952, it is the only proper, local score.



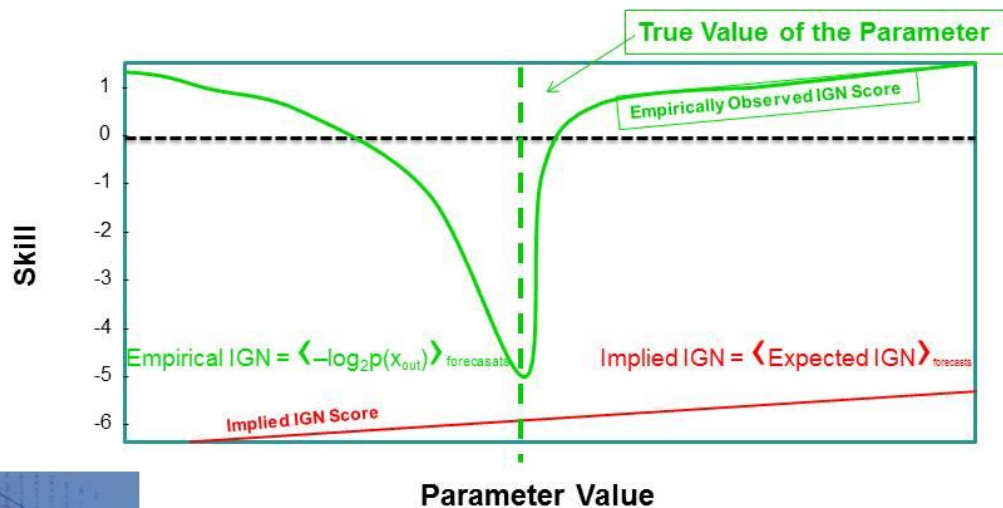
The parameter value isn't uncertain it is indeterminate. And so one way to avoid doing bad maths when doing physical science is don't make this mistake of thinking that something which is empirically vacuous is not defined. Don't think you are going to estimate it. You are looking for a value that does something, but there is no true value unless the "does something" is very precise. Great!

Identifying Parameters which Maximize Information (Minimize "IGN")

Given a Perfect Model Structure:

All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists), and the implied IGN reveals information deficit if the system is less than perfect.

$IGN = -\log_2(P(x_{outcome}))$ Proposed by I J Good in 1952, it is the only proper, local score.

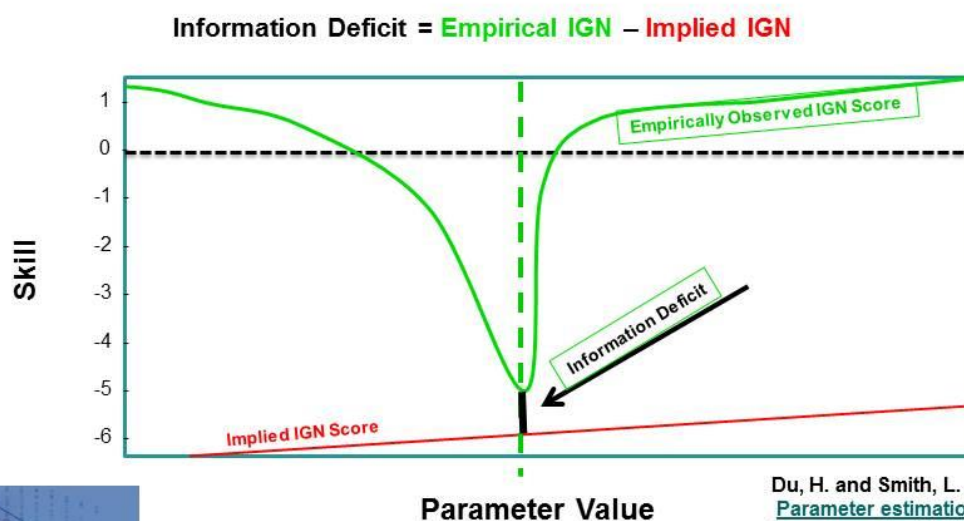


Identifying Parameters which Maximize Information (Minimize "IGN")

Given a Perfect Model Structure:

All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists), and the implied IGN reveals information deficit if the system is less than perfect.

$IGN = -\log_2(P(x_{outcome}))$ Proposed by I J Good in 1952, it is the only proper, local score.

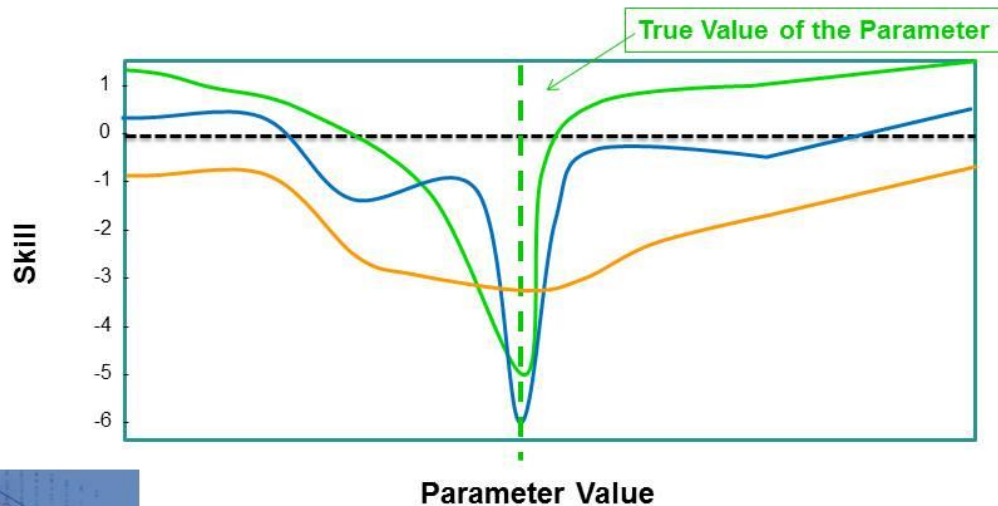


Uncertain parameters in perfect models



Given a **Perfect Model Structure**:

All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists). IGN has the bonus of providing the Info Deficit, as well as an easy interpretation in decision theory.



Indeterminate parameters in imperfect models

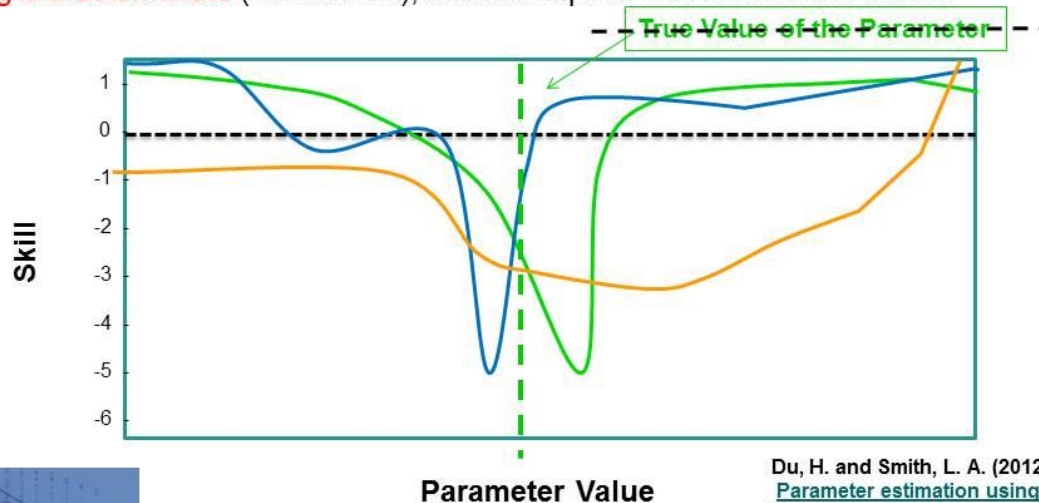


Given a **Perfect Model Structure**:

All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists).

Given an **Imperfect Model Structure**

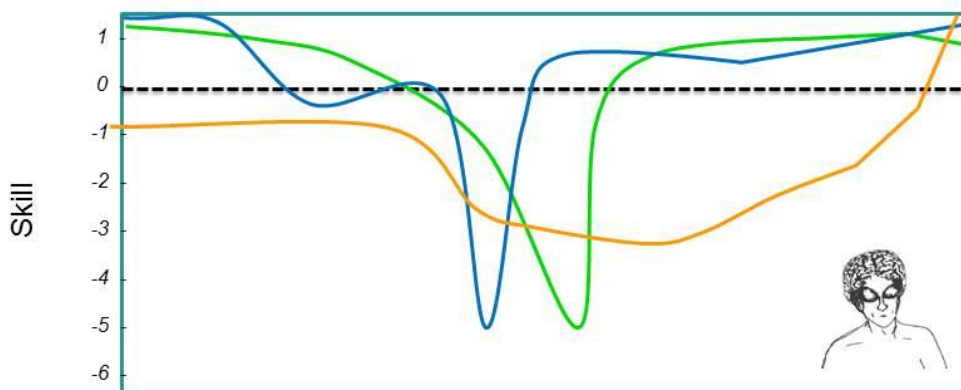
The particular Score matters, the forecast lead-time matters, **the value you are targeting is indeterminate** (none exists), and the implied IGN reveals info deficit.



Structural Model Error makes familiar tasks ill-posed

Given a Perfect Model Structure
There is no "True" value to identify
 All Proper Scores agree on the best parameter, the correct value is uncertain (but it exists).

Given an Imperfect Model Structure
 The particular Score matters, the forecast lead-time matters, the value we are targeting is indeterminate (none exists).



Du, H. and Smith, L. A. (2012) [Parameter estimation using Ignorance](#) PRE 86

Parameter Value



LA Smith, (2002) [What Might We Learn from Climate Forecasts?](#) Proc. National Acad. Sci. USA 4(99): 2487 -2492

LSE UUEM

Royal Society

Thur PM

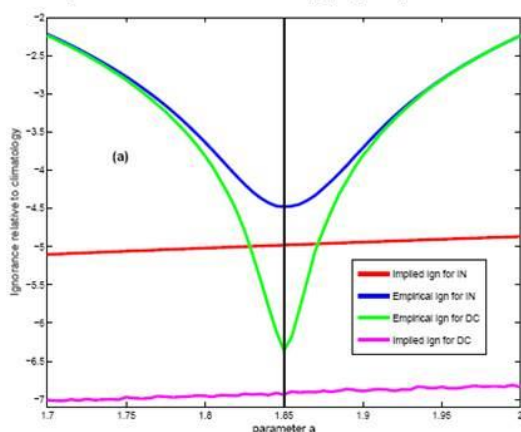
14 Sept 2017

Leonard Smith

Parameter Selection: Correct Model Structure

$$\text{Empirical IGN} = \langle -\log_2 p(x_{\text{out}}) \rangle_{\text{forecasts}}$$

$$\text{Implied IGN} = \langle \text{Expected IGN} \rangle_{\text{forecasts}}$$



Note that the Implied IGN
 $\langle \sum p(x) \log_2(p(x)/\mu(x)) \rangle$
 is less than the expected Empirical IGN
 $\langle \sum q(x) \log_2(p(x)/\mu(x)) \rangle$
 even at the correct value of a.

This Information Deficit(s) indicates that the (each) forecast scheme can still be improved.

Perfect Model Structure

All Proper Scores agree
 Data Assimilation Method Matters
 Target uncertain (but exists)
 Implied IGN reveals **information deficit**

H Du and LA Smith (2012) [Parameter estimation using ignorance](#) Physical Review E 86, 016213

Are weather forecasting models good enough that this does not matter?



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists)

Think Harder

Be as Transparent as you can be.

Find cool questions (useful and solvable) and serious people.

Have fun!

Accept Reality



LSE UUEM

Royal Society

Thur PM

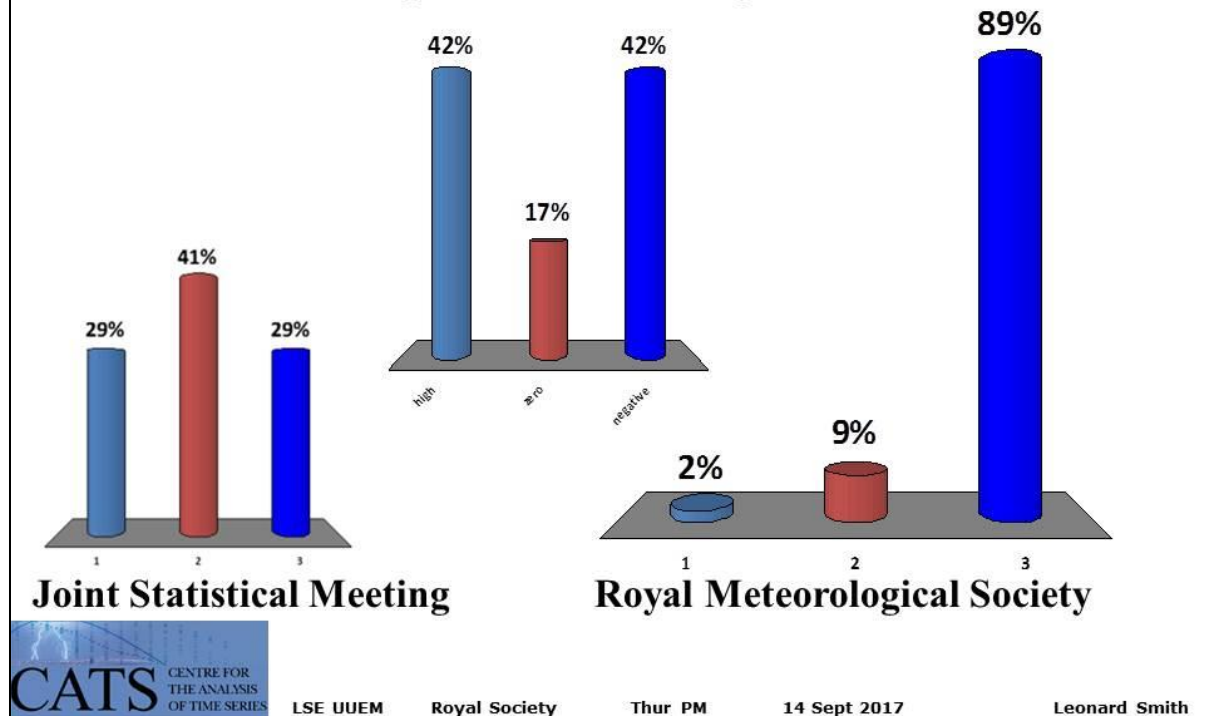
14 Sept 2017

Leonard Smith

So accept reality. This is reality. The models are wrong. All models are wrong, some are dangerous. Nevertheless you can make them... if you use them reasonably you can make them really a lot more useful. You want that one? I am not sure why she wants this one.

What does the user in the front line “want”?

“Captains of Industry”

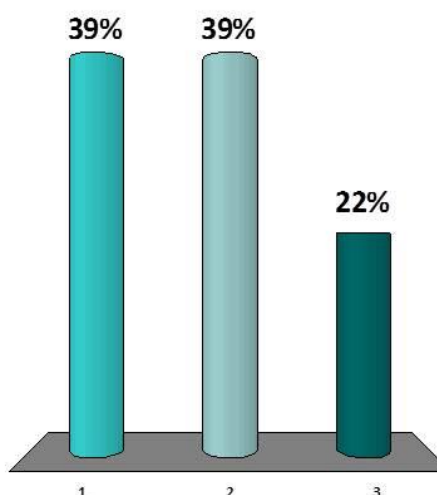


So what does a user want? The point is some of these guys really want a number, even if it is known to be wrong. Physical scientists don't want to give it to them. Statisticians don't care, or are things the other way around?

... regarding the model-based information you aim to create:

The utility of (known to be) unreliable quantitative information is:

1. High
2. Zero
3. Negative



This was you. You were being very stubborn. Ok? So this is a smaller sample than those others, but I think I am happy with that distribution. Maybe on different days, different days of the week I have these different views. So I am not going to talk about these guys. But there are other things that are very well accepted.

This Delphi method of doing it yourself and then doing it in smaller groups and larger. Coca Cola does this all over the world. They have groups in all these different places that make Coca Cola that talk, and then there are countries that get together and talk, and then the continents, they get together. Coca Cola is everywhere, right? They get together and talk and they all come together in Atlanta (not during a hurricane) and talk, and then they make a decision. And they find that they make decisions that way better. I think in that case I really believe it really works. Coca Cola also has tremendous climate statement, and they have been taking data since 1867 or something. But very often I see this. Even if you start adjusting for the number of people, which we haven't done on this graph yet. And so this idea that this Delphi method works, I mean just question everything. Maybe don't question everything - that is too harsh. Just ask for evidence that it is true in the particular setting that you are working on. That is just another one to say it really works. I think this is a perfectly fine answer. What were you trying to do? I don't know yet. You might want to see something work several times... you might want to see that this gives you insight. Someone has actually asked me - did you see this? Someone from ECMWF. One of the slides you tweeted must have had something about insight - "don't aim for statistics, aim for insight" or something like that.

So there was a tweet “Lenny what do mean ‘insight’?” I’m not sure yet, but I mean this is good. This is insight. If you think it is working and you can sort of get evidence that it really is working and the fact that you don’t understand yet means that maybe you should think harder. Think harder about what it is actually doing. And you are not alone. And what about the probabilities...these were just the other things in case it was really early. So what would I... also look for new insights under pressure, ok? I think there is something about having a deadline that is really good. Really good.

Focus on YOUR question!

The tools are not the product.

OBS Rule

Respect Newcomers

(Watchout for Jokers/Lobbyists/Fear)

Think Harder

Be as Transparent as you can be.

Find cool questions (useful and solvable) and serious people.

Have fun!

Accept Reality (maths and human)

Look for New Insights Under Pressure (67668)



LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

So this slide and all the slides that say focus on your question I wrote a slide in taxi 67668 on the way here from Goodenough College this afternoon. So I sort of had an idea of what I wanted to do. Erica will tell us what is interesting about this taxi number in a few seconds. But I mean there really is something nice about having to condense out what has happened over the three days here and yesterday in Reading in a hurry. And I think talking about it, just don’t be afraid to make mistakes. Try not to be afraid to make mistakes, or make really cool mistakes if you are going to make mistakes. But just don’t worry too much about it.



What is YOUR question?



13 Sept 2017 Predictability Probability and JEDI Insight University of Reading Leonard Smith

What is your question? I mean do you need more data if your question is what is the air speed velocity of a coconut-laden swallow? (Although I don't think that is a swallow.) The model isn't perfect but you still might get some evidence. Maybe all you really need is your question. Maybe if you have the right question that turns out to be enough, leads to other insightful things. Maybe you need to think outside the box. But just have fun and with that I wish you good luck, and hope you can really use some of what Erica has shown us over the last couple of days to make your work more fun and useful.



Thoughts?



Oxford Bus Shelter Sign:

X30+N30 predictions are wrong
sorry for any inconvenience

**Happy Trails with
YOUR Question!**



End

Background Reading:

- LA Smith (2002) *What might we learn from climate forecasts?* P. Nat. Acad. Sci (99)
LA Smith (2003) *Predictability Past Predictability Present. Predictability and Weather Forecasting* (ed. Tim Palmer, CUP).
LA Smith (2000) *Disentangling Uncertainty and Error*, in *Nonlinear Dynamics and Statistics* (ed A.Mees) Birkhauser.
Stainforth et al (2005) *Uncertainties in Prediction of Climate response*. Nature.
Stainforth et al (2007) *Uncertainty & Decision Support*. Phil Trans Roy. Soc. A, 1098

LA Smith (2007) *A Very Short Introduction to Chaos*. OUP
Nancy Cartwright (1983) *How the Laws of Physics Lie*. OUP



www.cccep.ac.uk
L.Smith@lse.ac.uk

TEMPUS PRODUCTIONS 25% VOTUM
SORRY FOR ANY INCONVENIENCE

When in doubt, distrusting the indications, or inferences from them (duly considered on purely scientific principles, and checked by experience), the words "Uncertain," or "Doubtful," may be used, without hesitation.

Fitzroy, 1862
Leonard Smith

CATS
CENTRE FOR
THE ANALYSIS
OF TIME SERIES

LSE UUEM

Royal Society

Thur PM

14 Sept 2017

Leonard Smith

Home | Help | Search | A-Z site index | LSE for You

You are here - Welcome to LSE > CATS > Publications

Publications http://www2.lse.ac.uk/CATS/publications/Publications_Smith.aspx

H Du and L A Smith (2012) '[Parameter estimation using ignorance](#)' Physical Review E 86, 016213
Smith, LA and Stern, N (2011) '[Uncertainty in science and its role in climate policy](#)' Phil. Trans. R. Soc. A (2011), 369, 1-24
R Hagedorn and LA Smith (2009) '[Communicating the value of probabilistic forecasts with weather roulette](#)'. Meteorological Applications 16 (2): 143-155. [Abstract](#)
Judd, CA Reynolds, LA Smith & TE Rosmond (2008) '[The Geometry of Model Error \(DRAFT\)](#)'. Journal of Atmospheric Sciences 65 (6), 1749-1772. [Abstract](#)
J Bröcker & LA Smith (2008) '[From Ensemble Forecasts to Predictive Distribution Functions](#)' Tellus A 60(4): 663.
J Bröcker, LA Smith (2007) '[Scoring Probabilistic Forecasts: The Importance of Being Proper](#)' Weather and Forecasting, 22 (2), 382-388. [Abstract](#)
J Bröcker & LA Smith (2007) '[Increasing the Reliability of Reliability Diagrams](#)'. Weather and Forecasting, 22(3), 651-661.
K Judd & LA Smith (2004) '[Indistinguishable States II: The Imperfect Model Scenario](#)'. Physica D 196: 224-242.
PE McSharry and LA Smith (2004) '[Consistent Nonlinear Dynamics: identifying model inadequacy](#)', Physica D 192: 1-22.
LA Smith (2003) '[Predictability Past Predictability Present](#)'. In 2002 ECMWF Seminar on Predictability. pg 219-242. ECMWF, Reading, UK. [Abstract](#)
MS Roulston & LA Smith (2002) '[Evaluating probabilistic forecasts using information theory](#)', Monthly Weather Review 130 6: 1653-1660. [Abstract](#)
LA Smith, (2002) '[What Might We Learn from Climate Forecasts?](#)' Proc. National Acad. Sci. USA 4 (99): 2487-2492.
D Orrell, LA Smith, T Palmer & J Barkmeijer (2001) '[Model Error in Weather Forecasting](#)', Nonlinear Processes in Geophysics 8: 357-371. [Abstract](#)
JA Hansen & LA Smith (2001) '[Probabilistic Noise Reduction](#)'. Tellus 53 A (5): 585-598.
LA Smith (2000) '[Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems](#)' (PDF) in Nonlinear Dynamics and Statistics, ed. Alistair I Mees, Boston: Birkhauser, 31-64. [Abstract](#)

CATS
CENTRE FOR
THE ANALYSIS
OF TIME SERIES

LSE UUEM Royal Society Thur PM 14 Sept 2017 Leonard Smith