Probabilistic Forecasting:
Why Model Imperfection Is a Poison Pill

Roman Frigg, Seamus Bradley, Reason L. Machete and Leonard A. Smith

Published in Hanne Anderson, Dennis Dieks, Gregory Wheeler, Wenceslao González and Thomas Uebel (eds): New Challenges to Philosophy of Science. Berlin and New York: Springer, 2013, 479-491.

1. Introduction

Foretelling the future is an age-old human desire, and the methods to pursue this goal are varied. Ancient Greeks consult an oracle; the superstitious ask a fortune teller to read the cards, and the rationally minded revert to scientific methods. Among the methods of science mathematical modelling has gained prominence: from planetary motion to nuclear fission, and from the growth of a population to the returns of an investment there is hardly a phenomenon that has not at one point or other been modelled mathematically. Many of these models make probabilistic forecasts: they provide us with probabilities for certain future events to occur. Weather models, climate models, financial market models, and hydrological models are but some prominent examples of models making probabilistic predictions. Designing such models is aided by the availability of ever increasing computational power, which has led to a trend of building ever larger and more complex models which are capable of making ever more precise predictions on an ever finer scale.

An example of the use of such a model is the recent project called *United Kingdom Climate Projections* (UKCP), which aims to make high resolution probability forecasts of the climate for up to 70 years from now. Figure 1 provides an example of such a forecast. It shows probabilities for different changes in precipitation under a medium level emission scenario.[1] The figure purports to tell us, for instance, that there is a 0.5 probability for a 20-30% reduction in precipitation in London by 2080. One of the striking aspects of this prediction is its precision. Calculations are made for a high resolution grid and so the forecast is able to distinguish, for instance, between the effects of climate change in London and Oxford (which are only an hour apart by train).

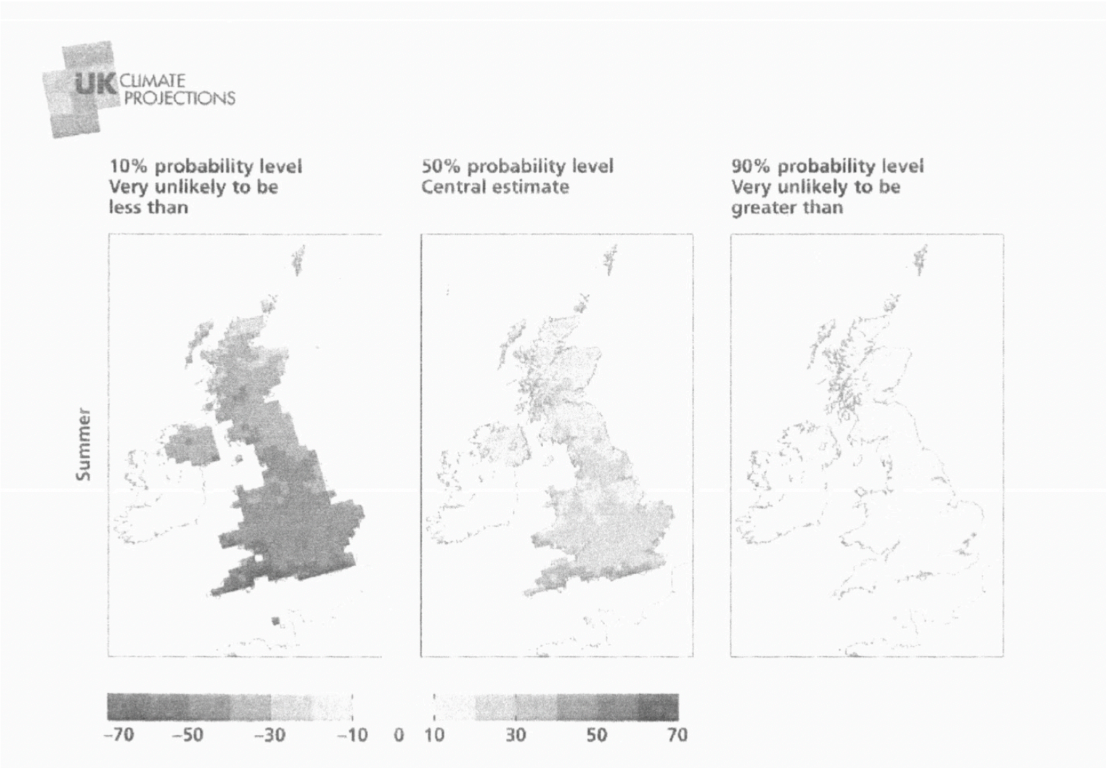

Figure 1: Change in summer mean precipitation (%) for the 2080s under a medium emission scenario. Source: UKCP.[2]

Computational power does not come for free. Super-computers are expensive tools, and developing and operating large computational models takes up the best part of the working hours of an ever increasing number of scientists. This raises the question of exactly what these models deliver: can these models provide the results as advertised?

---

[1] UKCP uses the IPCC A1B scenario. This is a kind of "optimistic" scenario of rapid growth and then a levelling off of the population by 2050 and a balance of renewable and fossil fuel energy. Total cumulative emissions amount to roughly twice what cumulative emissions were in 1990.

[2] http://www.ukcip.org.uk/wordpress/wp-content/UKCP09/Summ_Pmean_med_2080s.png; retrieved on 12 October 2011.

The aim of this paper is to urge some caution. We argue that if a model is non-linear and if there is only the slightest model imperfection, then treating model outputs as decision relevant probabilistic forecasts can be seriously misleading. This casts doubt on the trustworthiness of model results like the one we have just seen. In what follows we discuss this claim with a focus on climate modelling; we do so for the purpose of illustration and emphasise that the problem we describe crops up in all phenomena best modelled by non-linear models.

We begin by outlining the general methodology used in producing probabilistic forecasts, which we refer to as the *default position* (Section 2). Using computer simulations in a simple model we show that the default position produces seriously misleading results if the dynamics of the system is non-linear (Section 3). This casts serious doubt on the trustworthiness of model based probability distributions, and there is unfortunately no quick and easy way to dispel these doubts (Section 4). This raises serious questions about how models are (and indeed should be) used to make informed policy decisions (Section 5).

2. The Default Position

From a formal point of view, a climate model is a dynamical system, which we denote by (X, ϕ_t, μ). As the notation indicates, a dynamical system consists of three elements. The first element, X, is the system's *state space*, which contains all states in which the system could be. What these are depends on the nature of the system. For instance, the state space of a particle moving along a straight line are the real numbers, and the one of a hockey puck sliding on a square ice rink the unit square. The second element, ϕ_t, is the *flow* (or *time evolution*): if the system is in state $x \in X$ now, then it is in $y = \phi_t(x)$ at any later time t. In other words, ϕ_t, tells us how the system's state changes over the course of time. The third element, μ, is the system's Lebesgue measure: it allows us to say that parts of X have certain size. In case X is the real axis, μ is the length of an interval, and if X is the unit square the measure informs us what the area of parts of the square are. In the argument to follow μ plays no role and we note it here merely for the sake of completeness.

In the case of climate models X consists of relevant weather variables (such as air temperature, precipitation, wind speed, …), and ϕ_t tell us how they change over time. When descibed at that level of abstraction, one could be left under the impression that climate models are rather simple things. It is important to counter this impression before it gains traction. A full specification of the system's state space would involve giving the air temperature, precipitation, etc at *every point* on the surface of the earth! It is not only a practical impossibility to obtain these data; it is also an impossibility to store them with digital technology. For this reason we discretise the state space, meaning that we put a grid with a fininte number of cells on X and represent the state of an *entire cell* by one set of values for the relevant variables. The grid size is the length of the sides of the cells. Typically the grid size used in a climate model is well over 100km. Covering the world with such a grid still leaves us an enormous amount of data! Yet it is important to emphasise that the volume of numbers notwithstanding, this is a rather coarse description. For instance, the weather in the entire city of London is now represented by one set of numbers (one number for temperature, one for precipitation, etc.). The dynamics of the model raises even more issues. In order to specify ϕ_t we have to make a number of strongly idealising assumptions: we distort important aspects of the topography of the surface of the earth as the resolution of these models does not allow for realistic mountain ranges like the Andes, does not resolve the southern half of the state of Florida, many islands simply don't exist, including small volcanic islands chains easily visible in satelite photographs due to their interaction with clouds, and of course clouds fields themselves are not modelled realistically. Based on these idealising assumptions we can use basic physics (essentially fluid dynamics and thermodynamics) to formulate the equations of motion for the simplified earth's climate system. These equations are non-linear and we cannot solve them analytically. For this reason we resort to the most powerful computers available to compute solutions. The result of these computer simulations is ϕ_t.[3]

The formal apparatus developed so far has it that the flow takes as input a particular initial state x and then tells us into what state $y = \phi_t(x)$ this condition evolves under the dynamics of the system. Unfortunately this algorithm is not very useful in practice because we never know in what *exact* state the system is (if such a thing exists at all). To begin with, there is no measurement device that provides exactly correct values and so every measurement result comes with a certain margin of error. But more importantly, there is no such thing as *the* true wind speed in a model grid point corresponding to central London! Whatever number we settle on is an average of some kind or other; all we can truthfully say is something like 'the wind speed at a particular random location within that grid cell is likely to lie within a certain range'. We account for some of these uncertainties by specifying a probability distribution $p_0(x)$ over initial states, where the subscript indicates that the distribution describes our uncertainty about the initial condition at $t = 0$. There is of course a legitimate question about what the correct distribution is; we set this issue aside and assume that in one way or another we can come by the correct $p(x)$ (in the sense that it is a correct representation of our uncertainty).[4] The question then becomes: how does $p_0(x)$ change over the course of time? The flow ϕ_t can now be used to move $p_0(x)$ forward in time: $p_t(x) := \phi_t[p_0(x)]$.[5] This distribution is the central item of the *default position*,

---

[3] For a general introduction to climate modelling see Kendal McGuffie and Ann Henderson-Sellers, *A Climate Modelling Primer*. 3rd ed. New Jersey: Wiley 2005; a discussion of the specific model used in UKCP can be found at http://ukcp09.defra.gov.uk/.

[4] For a discussion of different kinds of uncertainty and their sources see Seamus Bradley, "Scientific Uncertainty: A User's Guide", in: *Grantham Institute on Climate Change Discussion Paper* 56, 2011 (available at http://www2.lse.ac.uk/GranthamInstitute/publications/WorkingPapers/Abstracts/50-59/scientific-uncertainty-users-guide.aspx).

[5] We use square brackets to indicate that $\phi_t[p_0(x)]$ is the propagating forward in time of the initial distribution $p_0(x)$. The flow of distribution derives from the flow of a state as follows: $p_t(x) := \phi_t[p_0(x)] = \Sigma_i p_0(z_i)$, where the sum of z_i reflects each of the states in X which are

the view that we obtain the decision-relevant probabilities for certain events to occur by plugging the initial distribution into the model and using the flow to obtain forecast probabilities for events at later times. The qualification 'decision-relevant' is crucial. The default position does not make the (trivial) statement that $p_t(x)$ is a probability distribution in a formal sense (i.e. that it is a mathematical object satisfying the axioms of probability); it is committed to the (non-trivial) claim that these probabilities are the correct probabilities for outcomes in the world in the sense that a rational decision maker should adjust his/her beliefs to these probabilities and act accordingly (assuming that there is no other pertinent evidence). In other words, $p_t(x)$ is taken to provide us with predictions about the future of sufficient quality that we ought to place bets, set insurance policies, or make public policy decisions according to the probablities given to us by $p_t(x)$.

3. The Poison Pill

Its intuitive appeal notwithstanding, the default position is wrong: $p_t(x)$ need not be the correct probability distribution, and taking $p_t(x)$ as a guide to actions can be ruinous.[6] Our strategy is to present a case where one can explicitly see that $p_t(x)$ need not be the correct probability distribution. This is enough to refute the default position, which has it that $p_t(x)$ *always* is the correct probability distribution.

Consider the following thought experiment. McMath has a pond in his garden where he breeds fish. He does not like being a hostage to fortune and wants to plan carefully how much food he will have to buy to feed his fish. To this end he constructs a model which allows him to predict the size of the population in his pond at a given time. He first introduces the population ratio ρ_t: the number of fish in the pond at time t

---

mapped onto x under the flow ϕ_t (i.e. $\phi_t(z_i) = x$ for all i); if the flow is invertible this reduces to $p_t(x) = p_0(\phi_{-t}(x))$.

[6] A probability distribution is deemed correct conditional on a particular observational resolution.

divided by the maximum number of fish the pond could accommodate; ρ_t lies in the unit interval $[0,1]$. To predict future populations he comes up with

$$\rho_{t+1} = 4\rho_t(1-\rho_t), \tag{1}$$

where time t is measured in units of weeks. So the model says that the population ratio in a week's time is four times today's ratio multiplied by one minus today's ratio.[7]

This allows McMath to predict the future size of his population given he knows today's size. The model is a dynamical system in the above sense with the unit interval $[0,1]$ being the state space, the flow being given by Equation 1, and the measure being the "usual" length of real intervals. So McMath decides to follow the prescription of the default position: he puts a probability distribution $p_0(x)$ over the initial conditions – here today's population ratio – and moves it forward in time under the dynamics of the system. He then uses the predictions thus generated to bet with one of his fellow villagers. The bet is "above or below 0.5": they split the unit interval into two equal parts, $[0,1/2]$ and $(1/2,1]$, which they call *A* and *B* respectively, and bet on whether *A* or *B* occurs in two months' time.

How successful will McMath be? Will he feed his fish well and will he win the bet against his mate? At this point the second part of our thought experiment begins: as we are pondering this question, we are incredibly lucky: heaven opens and God whispers the formula of the world's *true* dynamics into our ear:

$$\tilde{\rho}_{t+1} = 4\rho_t(1-\tilde{\rho}_t)\left[(1-\varepsilon)+\frac{4}{5}\varepsilon(\tilde{\rho}_t^{\,2}-\tilde{\rho}_t+1)\right], \tag{2}$$

[7] Equation (1) is of course just the well-known logistic map. The rationale for choosing this equation is that it is one of the simplest non-linear flows and that it has originally been proposed as a population model; see Robert May, "A Simple Mathematical Equation with very Complicated Dynamics", in: *Nature* 261, 1976, pp. 459-469. For the ease of presentation we assume that a new generation of fish is born once a week.

where ε is a parameter taken here to be 0.1. We immediately realise that this is just McMath's model plus a small perturbation. Figure 2 shows both the model (Equation 1) and the world (Equation 2), which makes obvious how similar the two are.

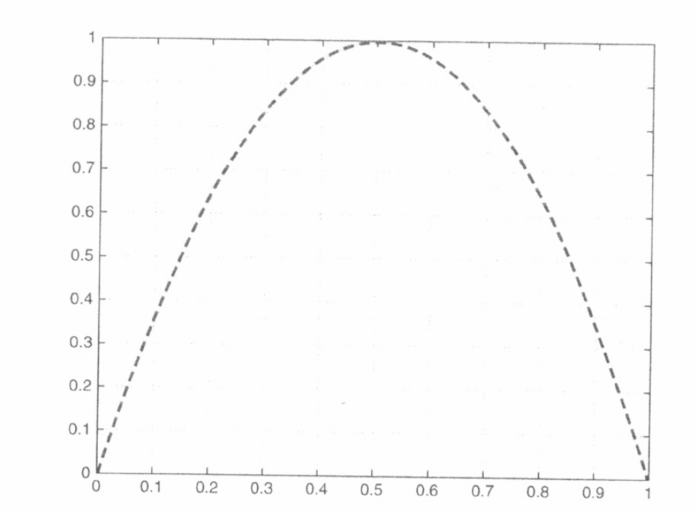

Figure 2: Equation 1 in blue (dotted line) and Equation 2 in yellow (drawn line) with ρ_t and $\tilde{\rho}_t$ on the x-axis and ρ_{t+1} and $\tilde{\rho}_{t+1}$ on the y-axis.

The maximum error of the model is 5×10^{-3} at $x = 0.85344$, where $\rho_{t+1} = 0.50031$ and $\tilde{\rho}_{t+1} = 0.49531$. This error is really small, which would lead us to believe that McMath's predictions should be accurate, and that therefore the use of the default position should be a winning strategy.

But calculations are better than intuitions, and so we use our God-given insider knowledge to see how well McMath will do. We move the initial distribution $p(x)$ forward in time both under the dynamics of the model (Equation 1) and the world (Equation 2), which gives us the two distributions $p_t^m(x)$ and $p_t^w(x)$ for the model and the world respectivley. If the default position was correct, one would have to find that $p_t^m(x)$ and $p_t^w(x)$ are identical, or at least broadly overlap. This is because, by assumption, the initial distribution is the correct distribution and the dynamics of the world is the true dynamics, hence $p_t^w(x)$ is the correct distribution and $p_t^m(x)$ captures what happens in the world only to the extent that it agrees with $p_t^w(x)$.

Since we don't know how to calculate $p_t^m(x)$ and $p_t^w(x)$ with pencil and paper, we resort to computer simulation. To this end, we divide the system's state space into 49 cells (which, in this context, are usually referred to as 'bins'). We then choose an initial distribution which is distributed according to the invariant measure within a radius of 7×10^{-3} from the true initial condition. The true initial condition is randomly chosen; in the concrete example to follow it happens to lie in the third bin. The true initial condition was within the same interval, but not necessarily at the centre. In turn we iterated forward in time 1024 points from the initial distribution (see first graph in Figure 3) under both dynamical laws. The other graphs in Figure 3 show how many points there are in each bin after two, fours and eight weeks respectively. Dividing these numbers by 1024 yields an estimate of the probability for the system's state to be in a particular bin.

[Insert Figures 3a-d so that they form a 2 by 2 matrix.
See the attach PDF for how the entire graph should look like]

Figure 3: The evolution of the initial probability distribution under the dynamics of the model (Equation 1, blue line) and the world (Equation 2, yellow line).

These calculations show the failure of the default position. While the two distributions overlap relatively well after two and four weeks, they are almost completely disjoint after two months. The implications of this for McMath are dramatic. His calculation led him to believe that after two months $p(A)\cong 1$ and $p(B)\cong 0$ (this is read off from the blue line in the fourth graph in Figure 3). This led him offer extremely long odds on *A*.[8]

[8] We use so-called *odds-for*, which give the ratio of payout to stake. These are convenient because they are reciprocals of probabilities; i.e. if $p(A)$ is the probability of A, then $o(A)=1/p(A)$ are the odds on A.

But the correct probabilities (read off from the pink line in the same figure) are $p(A) \cong 0.1$ and $p(B) \cong 0.9$. So he is very likely to lose a large amount of money to his fellow villager![9]

The moral of this thought experiment is that if a non-linear model differs from the truth only by a little bit (i.e. if the model has only a slight structural imperfection), then probabilistic predictions can break down. This implies that the default position is wrong. Simply moving forward in time an initial distribution under the dynamics of the model will not yield decision-relevant probabilities! But the break-down of the default position is nothing short of a methodological disaster: as we mentioned above, it is used in many places all the time and the realisation that probabilistic forecast cannot be trusted pulls the rug from underneath many modelling endeavours. One can sum up the result of our story in the slogan that model imperfection is a poison pill.

An immediate reply would point out that we have biased the presentation of the case in various ways to arrive at our conclusion and that the situation is in fact less dire than we make it out to be. The first bias is the focus on the two months forecast: had we focussed on the one month forecast McMath's forecasts would have been accurate enough to make both his planning and betting sustainable. Perhaps, perhaps not; but in the real world heaven doesn't open and no one whispers true dynamical laws into our ears. So we cannot simply compare the model with the true dynamics and affirm that we are fine at $t = 4$. In fact, if we knew the true dynamics we would not need a model in the first place! All we have is a model, and we know that the model is imperfect in various ways. What the above scenario shows is that model-probabilities and probabilities in the world can come unstuck dramatically, and as long as we have no means of telling *when* this happens, we better be on guard! For all we know there is no method of predicting when the model is accurate other than knowing the truth in advance, in which case we would not bother with a model anyway.

[9] Notice that our argument does not trade on worries about $p_0(x)$. We assume that the initial distribution gives us the correct probabilities and that setting ones degrees of belief in accordance with these probabilities would be rational. The core of our concern is what happens with these probabilities under the time evolution of the system.

The second alleged bias is the choice of the initial distribution. In order to run the calculations we have to choose a particular initial distribution (1024 points distributed according to the invariant measure within a radius of 7×10^{-3} from the true initial condition, which came to lie in the third bin). However, so the argument goes, this must be a special case that we have carefully chosen in order to drive home our sceptical conclusion; most of the distributions do not behave in this way and models provide trustworthy results most of time. Our story, so the counter continues, shows at most that every now and then unexpected results happen, but does not warrant a wholesale rejection of the default position.

There is no denying that our calculations rely on a particular initial distribution, but that realisation does not rehabilitate the default position. We repeated that same calculation with a large number of randomly chosen initial distributions, and it turns out that about one third of these distributions showed behaviour similar to that seen in Figure 3; another third resulted in forecast distributions that manifested an overlap of about fifty percent; and only one third behaved as the default position would lead us to expect. Hence, the problems we describe are by no means as rare as those critics would have it, and as long as we have no systematic way of drawing a line between the good and the bad distributions, we had better not rely too heavily on our calculations when making provisions for the future.

Some may have started wondering what all this has to do with modelling in the sciences; after all what we care about is the future climate or the stability of financial markets and not the fishing success of an imaginary Scotsman. Unfortunately the connection between our imaginary scenario and 'real' scientific cases is tighter and more immediate than we would like it to be. As we have seen above, the problems arise if models are non-linear and imperfect, and most scientific models have these properties. Without question, climate models have both these properties. It is not clear how to interpret the situation when different models agree (give indistinguishable probability forecasts), but in the climate case the different models give very different distributions (as becomes clear in the last IPCC Report, WG I) and so we know that the details of the models have a significant impact on expected results.[10] So when calculating, say, monthly precipitation in the 2080s based on climate models we may well not fare better with our planning of flood provision and water systems than McMath with fish food and bets.

4. Antidote Wanted

The first serious issue is whether Equations 1 and 2 are good proxies for all other non-linear systems. Equation 1 is of course the well-known logistic map with the independent parameter set equal to 4, which results in the dynamics being fully chaotic;[11] Equation 2 is a perturbed version of it. By saying that climate or finance models will face the same predictive breakdowns we implicitly assume that the problems when making predictions with the logistic map are typical of all non-linear models and will also occur in systems with completely different dynamical laws (as long as they are non-linear). It is fair to say that there is no hard and fast argument for this conclusion. However, it seems to us that the burden of proof lies with those who want to argue that the logistic map is a special case that the default position does not run into the problems we describe when used in the context of other non-linear models. Since the rise of chaos theory in the 1980s a bewildering array of non-linear systems has been studied and the general moral to be drawn from these studies is that random properties of systems get more dominant as (a) parameter values controlling the non-linear terms increase and (b) the size of the systems increases.[12] Generalising

---

[10] See Leonard Smith, "What Might We Learn from Climate Forecasts?", in: *Proceedings of the National Academy of Science* USA 4, 99, 2002, pp. 2487-2492.

[11] See Robert May, "A Simple Mathematical Equation with very Complicated Dynamics", *loc. cit.* and Leonard Smith, *Chaos. A Very Short Introduction*. Oxford: Oxford University Press 2007.

[12] By 'random properties' we mean, for instance, properties belonging to the ergodic hierarchy such as being mixing or Bernoulli; for a discussion of these see Joseph Berkovitz, Roman Frigg, and Fred Kronz, "The Ergodic Hierarchy, Randomness and Chaos", in: *Studies in History and Philosophy of Modern Physics* 37, 2006, pp. 661-691. An example of a system that becomes increasingly random as the perturbation parameter is turned up is the Hénon-Heiles system; see John Argyris, Gunter Faust, and Maria Haase, *An Exploration of Chaos*. Amsterdam: Elsevier 1994. For a discussion of systems that become more random as the number of particles increases see Roman Frigg and Charlotte Werndl,

from these cases one would expect that climate models, which are both strongly non-linear and huge, should display more rather than less of the problems we have seen above.

Another set of issues concerns lead times. Three challenges can be mounted. The first points out that all we are interested in are short term predictions and the above results show that in the short term the model forecasts appear accurate – hence there is no cause for concern. In some cases this seems to be the right response. In weather forecasting, for instance, we are mainly interested in predicting the immediate future and hence limiting model runs to the short term is the right thing to do. But this response does not seem to work in all cases. In both weather and climate modelling, for instance, we also are interested in the medium or long term behaviour and so we cannot limit predictions to short lead times. Of course what counts as short-term or long-term is relative to the model and it could be the case that by standards of the relevant climate models a prediction for 2080 is still a short term prediction. We are doubtful that this is the case. Indeed, it would be surprising to say if such predictions would turn out to be short term by the lights of a model used to make that prediction, in particular given that state of the art climate models differ even in terms of their performance over the past century. So we take it that the burden of proof lies squarely with those who believe that this is the case.

The second challenge argues for the opposite conclusion: what we are interested in is long term behaviour and so we can actually do away with detailed predictions completely and just study the invariant measure of the dynamics because it is the invariant measure that reflects a system's long term behaviour. Implicit in this proposal is the assumption that the invariant measures of similar dynamical laws are similar, because unless Equations 1 and 2 have similar invariant measures there is no reason to assume that adjusting beliefs according to the invariant measure is less misleading than adjusting them according to $p_t^m(x)$. However, it is at best unclear whether this is so. There is no proof that invariant measures have this property. Nonlinear systems are not expected to be structurally stable in general, which

---

Explaining Thermodynamic-Like Behaviour in Terms of Epsion-Ergodicity, in: *Philosophy of Science* 78, 3, 2011, pp. 628–652.

suggests that invariant measures need not be similar. And what is worse still, unlike McMath's pond, the world's climate is a transient system and as such it does not have an invariant measure at all.

The third challenge is that we are playing fast and loose with the notion of prediction. While McMath wants to predict what happens exactly two months from now, the above climate prediction is an average for the 2080s. So we would be comparing apples and pears. Not quite. What UKCP provides are not decadal averages. They provide an average for every year (the claim being that the distribution of likely monthly averages is the similar in each year of the 2080s). This is not so different from weekly predictions in the fish model. Other predictions made by UKCP are even more precise, e.g. the forecasts for the hottest day in August of a particular year. So what UKCP provides are not long term averages and hence an appeal to averages does not help circumventing the difficulties we describe.

5. Conclusion

We have argued that the combination of non-linear dynamics and model imperfection is a poison pill in that it shows that treating model outputs as probabilistic predictions can be seriously misleading. Probabilistic forecasts are therefore unreliable and do not provide a good guide for action.

This raises two questions. The first concerns the premises of the argument. The model being non-linear has been an essential ingredient of our story. While this assumption is realistic in that many relevant models have this property, there is still a question whether the effects we describe are limited to non-linear models. Arguably, if the world was governed by linear equations, then imperfect linear models need not suffer from the effects we discuss. One might like to avoid the assumption that the world is governed by any equations, of course, but the relevant point here is the role of model imperfections: a linear model will suffer from these effects unless its imperfections are also linear. The model being linear does not remove the difficulty we note. And of course, in practice the best models are not linear, nor are the relevant laws of physics.

The second question is what conclusion we are to draw from the insight into the unreliability of models. An extreme reaction would be to simply get rid of them. But this would probably amount to throwing out the baby with the bathwater because, as we have seen, in about one third of the cases the model indicates usefully. So the challenge is to find a way to use the model when it provides insight while guarding against damage when it does not. Finding a way of doing this is a challenge for future research.

Acknowledgments

We would like to thank audiences in Athens, Bristol, Ghent, London, Paris, and Toronto for valuable discussions.

Roman Frigg
Department of Philosophy
London School of Economics
Houghton Street
London WC2A 2AE
United Kingdom
r.p.frigg@lse.ac.uk

Seamus Bradley
Department of Philosophy
London School of Economics
Houghton Street
London WC2A 2AE
United Kingdom
s.c.bradley@lse.ac.uk

Reason L. Machete
Department of Mathematics and Statistics
University of Reading, P.O. Box 220
Reading RG6 6AX

United Kingdom

Leonard A. Smith

Centre for the Analysis of Time Series

London School of Economics

Houghton Street

London WC2A 2AE

United Kingdom

l.smith@lse.ac.uk