Model Error and Ensemble Forecasting: A Cautionary Tale

Seamus Bradley, Roman Frigg, Hailiang Du and Leonard A. Smith

1 Introduction

Many scientific enterprises nowadays involve using computer simulations to model complex phenomena of interest. Many models make probabilistic predictions using the methodology of *initial condition ensemble forecasting*, sometimes called ICE forecasting. Weather models, climate models, financial market models and hydrological models, among others, all do this sort of thing. We will see (using a simple example) how this a methodology successfully deals with certain kinds of uncertainty. The main aim of this paper is negative, however. We will show that ICE forecasting does not help with a certain other kind of errors, namely with structural model error. 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.

Models of the systems mentioned above are extremely complex. The problem we want to discuss shows up in much simpler models, so the remainder of our discussion will be about a particular simple model-system pair that we begin to introduce in the next section. After that, in Section 3 we introduce ICE forecasting and in Section 4 we demonstrate a problem with it. We relate our discussion of the simple model back to weather and climate models in Section 5 and we respond to some objections in Section 6.

2 Basic dynamical systems: the toy model

There is a pond in my garden that contains a population of fish. The fish population changes over time, and we would like to be able to predict those changes. So we are going to build a model of the pond. We describe the possible states of the system as a set X which we call the *state space*. This is a set that contains all the states the system (the pond) could be in. In the current example, the state space will be possible levels of fish population. Next we need to describe how the state of the system changes over time. We do this with a function we call ϕ_t which is the *time evolution* of the system. If the system is in state $x \in X$ now, the system will be in state $y = \phi_t(x)$ at a later time t. In other words, ϕ_t tells us how the system's state changes over time. So ϕ_t is a function

that, given as input the current state of the system, outputs the state of the system at some future time. These elements, X and ϕ_t, are two of the three ingredients of a *dynamical system*. The third element is a measure of the state space, μ. The measure μ assigns sizes to subsets of X.

We now want to look for a dynamical system that describes the population dynamics of the fish pond. To this end we introduce the population density ρ, the actual number of fish in the pond divided by the pond's maximum carrying capacity. So ρ_t is a number between 0 and 1. We adopt the following model [9]:

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

This equation gives us next week's fish population as a function of this week's population. The more fish there are this week, the more there will be next week (up to a certain point): hence the ρ_t term. As the population approaches one half of its maximum, this slows the growth of the population: hence the $1 - \rho_t$ term.

This equation is the well-known *logistic map*.[1] This is an example of a dynamics that demonstrates the property of *sensitive dependence on initial conditions*.[2] What this means is that fish populations that start off being very close together can diverge radically. That is, even if ρ_0 and ρ_0' are very close, it can be that ρ_t and ρ_t' are radically different for some later t. This means that such systems appear to become unpredictable, unless you know the *exact* initial conditions ρ_0. The next section discusses an interesting methodology used to overcome this problem.

3 Initial condition ensemble forecasting

Unfortunately we don't know the exact current fish population. We don't know what value of ρ_0 to plug in to get our prediction model going. We do have a rough idea of the current population, but we can't be sure exactly which is the "true" ρ_0. Given that the model exhibits sensitive dependence on initial conditions, just using our "best guess" of the initial condition guarantees no reliable prediction. It has become customary to use an ICE to deal with this problem (see, for instance, [1] and [5]). Let's say we take our best guess of the initial condition, and we centre some sort of "noise distribution" around this. The idea is that this distribution represents our uncertainty about the initial condition. What we want to do is to see how this distribution evolves through time.

So, instead of subjecting only our initial state x to the dynamics ϕ_t, we take some probability distribution around x, call it $p_0(x)$. Ideally we would subject that distribution to the dynamics. This gives us a new probability distribution $p_t(x) = \phi_t[p_0(x)]$. We shall leave aside issues relating to how one chooses p_0. Let's say that it is chosen somehow to best represent the state of uncertainty with respect to the variables in question. In

[1] For more on the logistic map see [9] and [10].

[2] For discussion of the concept of SDIC and its relation to predictability, see [10, Chapters 2, 9] and [11, Chapter 4].

our simulations we make the standard assumption that the distribution is a Gaussian distribution (truncated within the unit interval), but this assumption does no real work: similar results would obtain with other kinds of distributions. Our focus is on p_t. The implicit assumption that many make in working with p_t is that it represents a decision relevant probabilistic forecast of the target system. We call this the *default position*. Note that the default position does not make the trivial claim that p_t satisfies the mathematical axioms of probability. The default position is the stronger claim that the probability that this methodology produces can and should be used as the basis for decision making.

In practice we don't know how to subject the distribution to the dynamics analytically. So what we do is sample from the distribution, and subject each member of the sample to the dynamics. The initial conditions sampled from p_0 are called an ICE. Each element of the wnsemble is a possible initial condition for the system. We put every ensemble member through the dynamics and thereby create a collection of outputs. The set of outcomes of subjecting members of the ICE to the dynamics is called an ICE forecast. The distribution of outputs of this process is taken to represent the evolution of the distribution. In what follows, we sample 1024 ensemble members from a normal distribution centred on a particular initial condition, and we run each member through our dynamics. This distribution of ensemble members is then interpreted as the probability distribution of the future fish population.[3]

Is this a good methodology? Does this give us the right probability? First, it isn't clear what "the right probability" means: we are talking about a deterministic system, and the choice of noise distribution was somewhat arbitrary. So what does it mean to say that this method spits out the correct probability? Here's one way to assess whether the method is sound. Let's see how quickly agents who bet using their ensemble probabilities go bankrupt. We do the betting as follows. We split the state space up into a number of "bins",[4] and we use $p_t(x)$ to set the odds for the actual population ending up in that bin after t weeks. Then various bettors decide which bets to take. If the house's model probabilities are "good", then the bettors shouldn't be able to make huge sums of money off the house quickly and consistently.

The bettors in this scenario are betting simply according to what odds are being offered. That is, one agent will bet on an event if the inverse of the odds – the "implied-probability" – is in $(1/2, 1]$. Another bets on an event if its implied-probability is in $(1/4, 1/2]$, and so on.

Simulations show that these bettors don't make consistently huge gains. The house's model-based probabilities have therefore passed this test. So this is a promising methodology. The probabilities generated by the ICE forecast can usefully be taken as the basis for decision making.

[3]For further discussion of the interpretation of ensembles, see [5] and [4], and the other papers in that thematic issue of the Philosophical Transactions of the Royal Society.

[4]In our simulations there are 32 bins.

4 Model error

Sadly, it's not all good news. The spanner in the works is structural model error.[5] Let's imagine that the fish population dynamics is not given by the above model and that the actual dynamics are represented by the following equation:

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

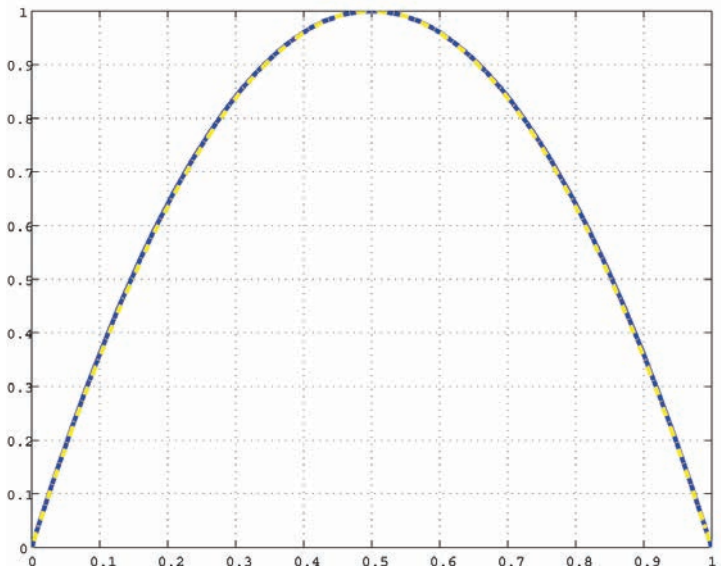

Figure 1: Graphs of Equation 1 (hatched line) and Equation 2 (drawn line).

As Figure 1 shows, the curves described by Equations 1 and 2 are very similar.[6] In fact, the maximum one-step error of the model is 5×10^{-3}, which is very small. One might expect that such a small error means that the model ensemble is still a pretty good guide to how to set your betting odds. Unfortunately, this is not the case.

We ran 1024 samples from our initial distribution through the dynamics of the system and through the dynamics of the model. We then produced histograms of number of ensemble members in each bin after 1,2,4 and 8 weeks.

As Figure 2 demonstrates, the system distribution and the model distribution can come apart radically in the middle term. In the short term, the distributions overlap, since the model error has not had time to have its effects blow up. In the long term, both distributions settle down into looking roughly similar (they have effectively converged on their invariant measures, which appear to be similar).[7] But in the middle term, the distributions look wildly different!

[5] We face structural model error when the model's functional form is relevantly different from that of the the true system.

[6] The figures in this paper are reproduced from [7].

[7] We have not shown figures of the long term, but by about $t = 16$ the distributions come to overlap again.

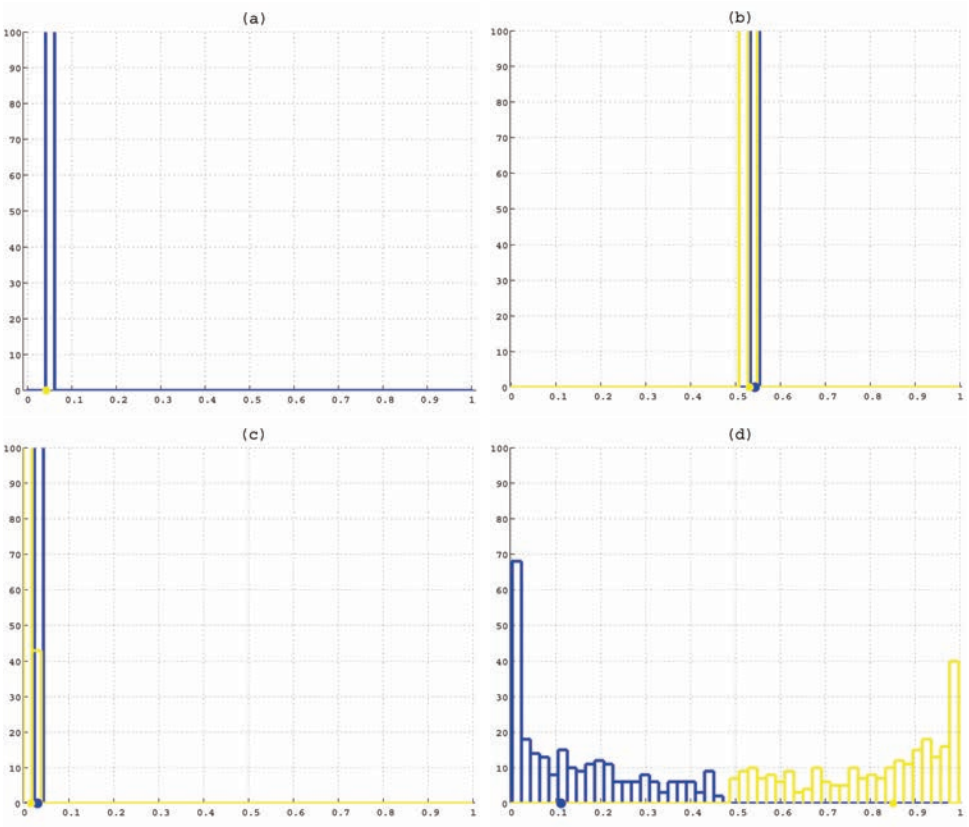

Figure 2: Histograms of ensemble members at lead times $t = 0, 2, 4, 8$.

It might be obvious that using the model-based probabilities as probabilities for decision-support would be unwise in this case, and probably disastrous. Consider the house's odds for the event "$\rho_8 \leq \frac{1}{2}$". That is, the event that the fish population will be less than half its maximum value after eight weeks. Almost all the model probability is in the left half of the $t = 8$ histogram. So the house sets short odds on this event, and very long odds on "$\rho_8 > \frac{1}{2}$". As the system probabilities show, however, almost all ensemble members actually land in the right half of the graph after 8 weeks. So it looks as if a house setting his odds on the model probabilities would lose a lot of money: it is putting very long odds on an event that is actually very likely to happen.

Simulations back this point. In the scenario discussed in [6], there are nine bettors. One will bet on an event if the inverse of the odds – the "implied-probability" – is in $(1/2, 1]$, another bets on an event if its implied-probability is in $(1/4, 1/2]$, and so on. They are simply betting according to what odds are being offered. They have no information that the house does not have. In other words, the information is completely symmetrical. The simulations show that the bettors mostly very quickly increase their wealth, thus the house must be losing money. But if betters who know no more than the house can rapidly increase their wealth, then there must be a serious problem with how odds are being offered.

One might push back against our worries by arguing that we have biased the discussion by a judicious choice of initial distribution. Sadly this is not the case. The calculations were repeated with 2047 different initial distributions. About a third of them behave in a way similar to the one seen in Figure 2; in about another third of the cases the two distribution came unstuck albeit in less dramatic fashion than in Figure 2; and in the last third of the cases the two distributions remained relatively close. So this problem is not one isolated to a particular initial distribution!

The conclusion is that offering odds corresponding precisely to model-based probabilities derived from a model with structural model error can have disastrous consequences and is therefore better avoided. There is a temptation to respond that this does not show that probabilities are useless; it only shows that we should not use these probabilities when they are misleading. The problem with this suggestion is that outside a thought experiment we have no means to tell when that happens. The only thing we have is the model which we know to be imperfect in various ways. We know that model-probabilities and probabilities in the world can come unstuck dramatically, but we don't know where and when. So while there is definitely information in the model-based probabilities it would be foolish to take them as actionable probabilities in decision support. As long as we have no means of of coping with the their shortcoming, wed better be on guard!

5 Weather and climate modelling

Recall that a dynamical system consists in a state space X which describes the possible states of the system, and a time evolution ϕ_t which describes how the system changes in time. In the case of a weather or climate model, the state space X will consist in a description of relevant atmospheric variables (air temperature, wind speed, humidity,

precipitation and so on). To fully describe the weather or climate system we would need to specify the values of all these variables at all points of the atmosphere. Such an endeavour is effectively impossible: first, even given the assumption that these quantities are well defined (effectively the continuum hypothesis), we don't have enough measuring devices to keep track of every variable at every location in the atmosphere, second we want to use computers to simulate the climate and the storage capacity of these devices is finite. We might therefore discretise the state space by putting a finite grid on it and describe the relevant variables in only a coarse-grained way. Even the smallest grid used in modern global climate models are about 100km on a side. This is a rather coarse description. For instance, the entire city of London is now represented by one set of numbers (one number for temperature, one for precipitation etc.). On top of this, the dynamics of the model introduces its own distortions. The ϕ_t for a climate model will involve many idealising assumptions. Important aspects of the topography of the surface of the Earth get distorted: the resolution of the models does not allow for realistic mountain ranges like the Andes; the Southern half of the state of Florida may be missing; many islands simply do not exist; clouds cannot be resolved properly.[8] Beyond merely failing to resolve certain processes directly, there is a deep question about what exactly is being represented in the model: there is no such thing as the true wind speed in the model grid point corresponding to central London.

Both weather and climate models thus suffer from various kinds of model error.[9] So the above cautionary tale should suggest that using ICE model-probabilities as decision relevant may be dangerous. And yet, it seems to be assumed by many people that such ensemble probabilities do provide decision-relevant information.

A recent example is a project entitled UKCP (United Kingdom Climate Projections), which aims to make high resolution probabilistic projections of the future climate for the United Kingdom for up to the end of the century.[10] It predicts, for instance, with a probability of 0.5 that there will be a 20–30% reduction in precipitation in London by 2080. These predictions are remarkably precise. The high resolution of the grid on which the predictions are made is able to distinguish the effects of climate change in London and Oxford (which are only an hour apart by train).

Developing and operating these large computational models is a significant cost to the groups who do it, and an increasing number of scientists are devoting their careers to building and running such models. The question naturally arises whether these models can deliver as advertised. The insights gained in the last section would urge both some caution in decision making and some care in resource allocation in the modelling process.

[8] There will of course be parameters that attempt to simulate the effects of things like clouds.

[9] For more on uncertainty in science, see [3].

[10] UKCP's probabilities are derived in a more complicated way than the ones discussed in the last section, but the issues we describe equally apply to more complex schemes. For a discussion of UKCP's methodology see [8].

6 Ways out?

In this section we discuss various ways one might try to avoid our negative conclusions. First, we return to the question of whether we are interested in the "middle term" – where the distributions come apart drastically. Some have objected to our argument by saying that we are in fact only interested in short term predictions and thus our results do not pose a problem.

Of course it we may sometimes be interested in the short term, but in both weather and climate modelling we also are interested in the medium or long term behaviour: we do not limit predictions to short lead times. 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 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.

Going the other way, some critics argue that one only ought to be interested in the long term behaviour of the system. Such behaviour is dominated by the structure of the invariant measure, and so all we need to do is study the invariant measure. Implicit in this proposal is the assumption that the invariant measures of the system and the model are similar. While this assumption may be true for the toy example discussed above, it is not true in general. First, we have no reason to assume that the non-linear systems we are interested in are *structurally stable* in the sense that similar systems have similar invariant measures. Second, the actual system is a transient system (that is, when the climate is changing) and as such does not have an invariant measure at all.

A further possible defence of the default position might just take our results to be inapplicable to actual modelling practices because wether and climate models are very different from our simple logistic map case. We agree that the models are very different, but as we saw in the last section, weather and climate models are likely to have even more serious model errors than in our toy example. We think the burden of proof is on the proponent of the default position to show what it is that makes such models immune from our criticism. The weather and climate modelling methodology is obviously much more complex than the simple case we presented above, but are any of the relevant differences such as to invalidate the applicability of our conclusions? We don't think so.

It is fair to say that there is no hard and fast argument for this conclusion. It seems to us, however, 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.

References

[1] Anderson, Jeffrey L. Selection of initial conditions for ensemble forecasts in a simple perfect model framework. Journal of the Atmospheric Sciences 53(1), 1996, 22-36.

[2] Argyris, John, Gunter Faust and Maria Haase. *An exploration of chaos*. Elsevier, 1994.

[3] Bradley, Seamus. Scientific uncertainty: A user's guide. Technical Report 56, Grantham Institute Working Paper, 2011.

[4] Collins, Matt Ensembles and probabilities: a new era in the prediction of climate change. *Philsophical Transactions of the Royal Society*, 365:1957–1970, 2007.

[5] Leutbecher, Martin and Tim N. Palmer Ensemble forecasting Journal of Computational Physics 227 (2009) 3515-3539.

[6] Frigg, Roman, Seamus Bradley, Hailiang Du, and Leonard A. Smith. Laplace's demon and climate change. Technical Report 103, Grantham Institute Working Paper, 2013.

[7] Frigg, Roman, Seamus Bradley, Reason L. Machete and Leonard A. Smith. Probabilistic forecasting: why model Imperfection is a poison pill forthcoming in Hanne Anderson, Dennis Dieks, Gregory Wheeler, Wenceslao Gonzalez and Thomas Uebel (eds): New Challenges to Philosophy of Science. Berlin and New York: Springer

[8] Frigg, Roman, David Stainforth, and Leonard A. Smith. The myopia of imperfect climate models: the case of UKCP09. *Philosophy of Science*, 2013.

[9] May, Robert Simple mathematical models with very complex dynamics. *Nature*, 261:459–467, 1976.

[10] Smith, Leonard A. *Chaos: A very short introduction*. Oxford University Press, 2007.

[11] Smith, Peter *Explaining chaos*. Cambridge University Press, 1998.