9

Predictability past, predictability present

Leonard A. Smith
Centre for the Analysis of Time Series,
London School of Economics and Pembroke College, Oxford

Maybe we oughta help him see,
The future ain't what it used to be.
Tom Petty

9.1 Introduction

Predictability evolves. The relation between our models and reality is one of similarity, not identity, and predictability takes form only within the context of our models. Thus predictability is a function of our understanding, our technology and our dedication to the task. The imperfection of our models implies that theoretical limits to predictability in the present may be surpassed; they need not limit predictability in the future. How then are we to exploit probabilistic forecasts extracted from our models, along with observations of the single realisation corresponding to each forecast, to improve the structure and formulation of our models? Can we exploit observations as one agent of a natural selection and happily allow our understanding to evolve without any ultimate goal, giving up the common vision of slowly approaching the Perfect Model? This chapter addresses these questions in a rather applied manner, and it adds a fourth: Might the mirage of a Perfect Model actually impede model improvement?

Given a mathematical dynamical system, a measurement function that translates between states of this system and observations, and knowledge of the statistical

Predictability of Weather and Climate, ed. Tim Palmer and Renate Hagedorn. Published by Cambridge University Press.

characteristics of any observational noise, then in principle we can quantify predictability quite accurately. But this situation describes the perfect model scenario (PMS), not the real world. In the real world we define the predictability of physical systems through our mathematical theories and our *in silico* models. And all our models are wrong: useful but imperfect. This chapter aims to illustrate the utility of existing ensemble prediction systems, not their imperfections. We will see that economic illustrations are of particular value, and investigate the construction of probability forecasts of observables from model simulations. General arguments and brief illustrations are given below; mathematical details and supporting statistics can be found in the references. While the arguments below are often couched in terms of economic users, their implications extend to the ways and means of meteorology as a physical science. Just as it is important not to confuse utility with empirical adequacy, it is also important to accept both as means of advancing any physical science.

In the next three sections we make a quick tour of useful background issues in forecasting, economics and predictability. When considering socioeconomic value it is helpful not to confuse severe weather and high-impact weather: the value of a weather forecast depends not only on its information content but also on our ability to take some mitigating action; a great deal of the unclaimed value in current operational products lies in their ability to yield useful information regarding unremarkable weather which carries significant economic impact. Then in Section 9.5 we consider the question of comparing forecasts and the notion of 'best'. This continues in Section 9.6 with a number of issues at the interface of meteorology and statistics, while illustrations of their economic relevance are noted in Section 9.7. It is quite popular nowadays to blame forecast busts on 'uncertainty in initial condition' (or chaos), we discuss what this phrase might possibly mean in Section 9.8, before concluding in Section 9.9. In reality predictability evolves and, as shown in Section 9.4, 'the future' evolves even within the mathematical fiction of a perfect model scenario where predictability does not.

9.2 Contrasting 1995 and 2002 perspectives on predictability

What has changed in the short time since the 1995 ECMWF Seminar on Predictability? Since I cannot avoid directly criticising what was happening in 1995, we will focus mostly on my contribution to the seminar (Smith, 1996).

Ensemble formation for systems of chaotic differential equations was a major topic of discussion in 1995; in contrast this chapter does not contain a single differential equation. In fact, it contains only one equation and, as it turns out, that equation is ill posed. In 1995 my focus was on constructing perfect ensembles, while below we will be more concerned with interpreting operational ensembles. The 1995 paper quantifies the difference between some forecast probability density function (pdf)

and a perfect pdf obtained by propagating current uncertainty forward in time under a perfect model, while below I am content to discuss how to change an ensemble of simulations into a pdf forecast in the first place. There is also a question as to how one should evaluate any forecast pdf, given that we never have access to the 'perfect pdf', if such a thing exists, but only observations of a single realisation of weather. That is, we have only measurements of the one thing that happened, a target often called the *verification*. In general, it seems to me that the 1995 discussion focused on doing maths within the perfect model scenario (PMS), whereas we are now more interested in quantifying information content and debating resource allocation.

Smith (1996) discussed quantifying model error, while now I have been reduced to pondering model error, which I now refer to as *model inadequacy* (following Kennedy and O'Hagan, 2001). Any particular model can be thought of as one member of a model class. As a very simple example, consider different models that share the same structural form but have different parameter values, these are in the same model class; or consider the collection of all one-dimensional maps consisting of finite-order polynomials. Model inadequacy reflects the fact that not only is the best model we have imperfect, but there is no member of the available model class which is perfect. This is a much deeper flaw than having incorrect parameter values: in this case there are no 'Correct' parameter values to be had. And this case is ubiquitous within physical science.

The concept of *i*-shadowing was introduced in the 1995 Predictability Seminar, as was the notion of an accountable probability forecast. A model is said to *i-shadow* over a given period in time if there exists a model trajectory that is consistent with the observations, given the observational noise, over that period. For historical reasons, meteorologists call the model state that corresponds to the operational best guess of current atmospheric conditions *the analysis*. The question of quantifying just how long operational models can shadow either the observations or even the corresponding time series of analyses remains of key interest. The notion of an accountable ensemble forecast was also introduced in the 1995 Seminar (see also Smith, 2001) as a generalisation of Popper's idea of accountability in the single forecast scenario. Popper (1956) realised that forecasts would fail due to uncertainty in the initial condition even if the model was perfect; he called a model accountable if it correctly specified the accuracy of measurement required to obtain a given forecast accuracy. For an accountable ensemble forecast the size of the ensemble will accurately reflect the resolution of the probability forecast. The relevant point here is that any forecast product extracted from an accountable probability forecast will suffer *only* from the fact that the forecast ensemble has a finite number of members: we could never reject the null hypothesis that both the members of the forecast ensemble and the verification ('Truth') were drawn from the same distribution.

The distribution of shadowing times is arguably the best measure we have for contrasting various non-linear models and quantifying the relevance of uncertainty in the initial condition (as opposed to model inadequacy). I hope that in the pages

that follow methods which reflect the quality of a simulation model (i.e. shadowing times) are clearly distinguished from methods which reflect the quality of a complete probabilistic forecast system (i.e. ignorance as defined in Good (1952) and Section 9.5.5). Recall that with a non-linear model, a probabilistic forecasting system can only be evaluated as a whole: non-linearity links data assimilation, ensemble formation, model structure and the rest.

In general, I would identify two major changes in my own work from 1995 to 2002. The first is a shift from doing mathematics to doing physics; more specifically, of trying to identify when very interesting mathematics is taking us to a level of detail that cannot be justified given the limited ability of our model to reflect the phenomena we are modelling. Indeed, I now believe that model inadequacy prevents accountable probability forecasts in a manner not dissimilar to that in which uncertainty in the initial condition precludes accurate best first guess (BFG) forecasting in the root-mean-square sense. In fact we may need to replace our current rather naïve concept of probability forecasts with something else; something which remains empirically relevant when no perfect model is in hand. The second change reflects a better understanding of the role of the forecast user as the true driver for real-time weather forecasting. Economic users can play particularly vital roles both as providers of valid empirical targets, the ultimate test of mathematical modelling (at least within mathematical physics), and also as a valuable source of data for assimilation. In the next section, we will develop an ensemble of users with which to illustrate this interaction.

9.3 An ensemble of users

Tim Palmer's chapter in this volume (see also Palmer, 2002) introduced his golf buddy, Charlie the contractor. Charlie is forced by the nature of his work to make binary decisions, for example whether or not to pour concrete on a given afternoon. The weather connection comes in as another binary event: if it freezes then the cement will not set properly. By using cost-loss analysis (Angstrom, 1919; Murphy, 1977; Richardson, 2000), Charlie can work out the probability threshold for freezing at which he should take the afternoon off and go play golf. Of course, if Charlie is presented with a definitive forecast ('The low temperature tonight will be 4 degrees C' or 'No ground frost tonight'), then pours the concrete and it does freeze, he is likely to be somewhat disappointed. As Tim noted, he may look for someone to sue. There are, of course, no definitive forecasts and to sell any forecast as unequivocal is to invite lawsuits. When a forecaster has foreknowledge of the uncertainty of a forecast and yet still presents an unequivocal forecast to the public, justifying it as being 'for their own good', she is inviting such a law suit. Arguably, the public has the right to expect a frank appraisal of the forecaster's belief in the forecast. As it turns out, Charlie also plays the horses; he knows much about odds. He does not

even hold the naïve expectation that the corresponding implied 'probabilities' (of each horse winning) should sum up to one!

But there is more to the world than binary decisions (and golf). While I do not know Charlie, at a recent London School of Economics alumni dinner I met Charles. Charles now works in the financial futures market; while he no doubt plays golf, he does not see himself as making binary decisions; he is interested in 'How much' questions rather than 'Yes or No' questions. This is because Charles buys and sells large quantities of petroleum products (heating oil, gas, various flavours of crude, jet fuel and so on) always being careful not to take delivery of any of it. He has also started pricing weather derivatives, a wide variety of weather derivatives in fact. He is fluent in stochastic calculus and knows a bit of probability theory, enough to know that in order to gauge his risk he wants more than a single probability threshold.

Charles has an interesting view of what constitutes a good four-week forecast. He doesn't care at all about the average temperature in week four, nor whether the Monday two weeks hence is in fact going to be a very cold day. While Charlie is concerned as to whether or not cement will set tonight, Charles is not concerned about any particular day. Instead, Charles is very concerned about the number of days between now and the end of the month on which cement will not set, inasmuch as this is the kind of variable that weather derivatives near expiry might be based upon. Charles knows how to place his bets given a good probability forecast; his question is whether a given probability forecast is a good one! For better or for worse, one of the advantages of providing a probability forecast is that no single probability forecast need ever be judged 'wrong'; finding oneself overly worried about what might happen in any one forecast suggests we have not fully accepted this basic lesson of first-year Statistics. Nevertheless, if Charles only bets when the forecast probability of winning is over 90% and, after many bets, he finds he has won only half the time, then he will have a strong case against the forecast vendor.

The financial markets are inundated with vendors of various forecasts, and Charles is familiar with forecasts that fail to provide value. He already knows how to sue, of course, but never does so; life is too short. Rather, he relies on natural selection in the marketplace: if the forecasts do not contain the information he needs, or are not presented in a manner such that he can extract that information (even if it is there), then he simply stops buying them and speaks badly of them in London wine bars.

And then there is Charlotte, another LSE graduate now working in the energy sector. Charlotte's goal is not to make money *per se*, but rather to generate electricity as efficiently as possible, using a mix of wind power alongside a set of combined cycle gas turbine (CCGT) generators.[1] Her definition of 'efficiently' is an economic one, but includes issues of improving air quality and decreasing carbon dioxide production. Ideally, she wants a forecast which includes the full probability density function (pdf) of future atmospheric states, or at least the pdf for temperature, pressure, wind speed and humidity at a few dozen points on the surface of the Earth. And she would like to forecast weather dependence of demand as well, especially where it is sensitive

to variables that also impact the efficiency of generation (Altalo and Smith, 2004; Smith *et al.*, 2005). The alternative of using only marginal distributions, or perhaps a bounding box (Smith, 2001; Weisheimer *et al.*, 2005), may prove an operational alternative. Charlotte's aim is not to be perfect but rather simply to do better, so she is happy to focus on a single site if that is required by the limited information content in the forecast.

A quick calculation[2] shows that even with an accountable forecast the ensemble size required in order to resolve conditional probabilities will remain prohibitive for quite some time. Of course, the future may allow flow-dependent resource allocation, including the distribution of ensemble members over multiple computer sites on those days when such resources are justified. A somewhat longer and rather more dubious calculation suggests that generating this style of weather forecasting might feed back on the weather which is being forecast. Certainly the effect would far exceed that of the flapping of a seagull's wing, unless the forecasters were relocated to some remote location, say on the moon.

Like Charles, Charlotte is also interested in the number of cold days in the remainder of this month, or this season. This is especially true if they are likely to be consecutive cold days. The consumption of natural gas, and hence its price, depends on such things. And it is impossible to deduce them from knowing that there is a 10% chance of a cold day on four consecutive days in week two: thinking of the ensemble members as scenarios, Charlotte wants to know whether that 10% is composed of the same ensemble members each day (in which case there is a 10% chance of a four-day cold spell) or whether it was the case that 10% of the total ensemble was cold on each of the four days, but the cold temperatures corresponded to different members each day, in which case the chance of an extended cold spell could be very low (depending on the size of the ensemble).

Why does she care about the four consecutive cold days problem? By law (of Parliament), natural gas will be diverted from industrial users to domestic users in times of high demand. If she can see that there is a moderate probability of such a period in advance, she can fill her reserve tanks before the start of the cold spell (and take a forward position in gas and electricity markets as well). This is a fairly low cost action, because if the cold spell fails to materialise, she can simply decrease purchase of natural gas next week. The carrying costs of an early purchase are small, the profit loss of running low is huge: she is happy to overstock several times a year, in order to have full reserves during the cold spells that do occur.

And she can mix this weather information in with a variety of other indicators and actions; from scheduling (or postponing) preventive maintenance, to allowing optional leave of absence, so as to embed probabilistic weather information naturally into a scheme of seamless forward planning.

Decisions similar to Charlotte's are made across the economy. We consider two more examples of such decisions: one faced by the owner of a small corner shop and one faced by a multinational energy giant. By overstocking soft drinks whenever

the probability of a heat wave exceeds a relatively low threshold, shop owners can hedge against the significant impact of running out of stock when a heat wave finally materialises. The marginal costs of the extra stock are relatively low, as in Charlotte's case, and the shop owner will happily make this investment several times when no heat wave materialises. We find something like the inverse of the understock problem in offshore oil wells. In deep water, floating oil wells store the extracted oil onsite, until a tanker can come and collect it; for a variety of reasons, it is costly for the local storage tanks to get too full. To offload the oil, a tanker must not only reach the well, but seas must be calm enough for it to dock; obviously it can make sense to send a tanker early if there is a good chance of sustained heavy seas in the vicinity of the well near the originally scheduled time of arrival.

Cost functions targeting Charles' and Charlotte's desires could prove very valuable to modellers; inasmuch as we have not already fit our models to such targets, they provide a fresh viewpoint from which we can detect hidden shortcomings of our models. But even beyond this, Charlotte and her colleagues are not only collecting traditional meteorological observations at various points scattered about the country (power station locations), they also collect real-time data on weather-related demand integrated over spatial regions comparable with the grid resolution of a weather model and on timescales of seconds: the assimilation of such observations might also prove of value. This value will hinge on the information content of the observations, not their number: a huge number of uninformative or redundant observations may tell us much less than a few relevant measurements.

Charlie, Charles and Charlotte each aim to extract as much relevant information as possible from the forecast, but no more. Each of them realises that, in the past, weather forecasts have been presented as if they contained much more information than even a casual verification analysis would support. The five-day forecasts for Oxford at www.metoffice.com present the day 5 forecast with the same air of authority given to the day 2 forecast. The Weather Channel, which provides 10-day point forecasts at www.weather.com, and other vendors are concerned to present the uncertainty they know is associated with their current apparently unequivocal forecasts. Yet a decade after ensemble forecasting became operational, it is still not clear how to do so. And the situation is getting worse: there is talk of commercially available point forecasts out to 364 days, each lead time presented as if it were as reliable as any other (in this last case, no doubt, the vast majority are equally reliable). Questions of how best to communicate uncertainty information to numerate users and how to rapidly communicate forecast uncertainty to the general public are central to programmes like THORPEX. Answering these questions will require improving the research interface between social psychology, meteorology and mathematics. Our current progress in this direction can be observed in the forecasts posted at www.dime.lse.ac.uk (see also http:// lsecats.org and www.meteo.psu.edu/~roulston).

So we now have our ensemble of three users, each with similar but distinct interests in weather forecasts and different resources available to evaluate forecast information.

Charlie is primarily interested in binary choices. Charles' main interests lie both in a handful of meteorological standards (where he only cares about the legal value of the standard, not what the weather was) and in very broad meteorologically influenced demand levels. Charlotte is interested in accurate, empirically falsifiable forecasts; she doesn't care what the analysis was nor what the official temperature at Heathrow was. She has temperature records of her own in addition to measurements of the efficiency of her CCGT generators and wind turbines. In a very real sense, her 'economic variables' are more 'physical' than any model-laden variable intended to reflect the 500 mb pressure height field within some model's analysis.

There are, of course, important societal uses of weather forecasts beyond economics; many of these societal applications are complicated by the fact that the psychological reactions figure into the effectiveness of the forecast (for an example, see Roulston and Smith, 2004). For most of what follows, however, we will consider weather forecasts from the varying viewpoints of Charlie, Charles and Charlotte. Obviously, I aim to illustrate how probabilistic forecasts derived from operational ensemble prediction systems (EPS) compare, in terms of economic relevance, with forecasts derived from a best first guess (BFG) approach, and we shall see that probabilistic forecasts are more physically relevant as well. But I will also argue that accountable probability forecasts may well lie forever beyond our grasp, and that we must be careful not to mislead our 'users' or ourselves in this respect. To motivate this argument, we will first contrast the Laplacian view of predictability with a twenty-first century view that accounts for uncertainty in the initial condition, if not model inadequacy.

9.4 Contrasting nineteenth vs. twenty-first century perspectives on predictability

Imagine (for a moment) an intelligence that knew the True Laws of Nature and had accurate but not exact observations of a chaotic system extending over an arbitrarily long time. Such an agent, even if sufficiently powerful to subject all this data to (exact) mathematical analysis, could not determine the current state of the system and thus the present, as well as the future, would remain uncertain in her eyes. Yet the future would hold no surprises for her, she could make accountable probability forecasts, and low probability events would occur only as frequently as expected. The degree of perfection that meteorologists have been able to give ensemble weather forecasting reflects their aim to approximate the intelligence we have just imagined, although we will forever remain infinitely remote from such intelligence.[3]

It is important to distinguish determinism and predictability (see Earman, 1986; Bishop, 2003 and the references therein). Using the notion of indistinguishable states (Judd and Smith, 2001, 2004) we can illustrate this distinction with our twenty-first century demon, which has a perfect model and infinite computational capacity but

access to only finite resolution observations. If the model is chaotic, then we can prove that, in addition to the 'True' state, there exists a set of states that are indistinguishable from the particular trajectory that actually generated the data. More precisely, we have shown that given many realisations of the observational noise, there is not one but, in each case, a collection of trajectories that cannot be distinguished from the 'True' trajectory given only the observations. The system is deterministic, the future trajectory of each state is well defined and unique, but uncertainty in the initial condition limits even the demon's prediction to the provision of probability distributions.

Note that the notion of shadowing is distinct from that of indistinguishable states (or indistinguishable trajectories); *i*-shadowing contrasts a trajectory of our mathematical model with a time series of targets usually based on observations of some physical system. This is often cast as a question of existence: does the model admit one or more trajectories that are consistent with the time series of observations given the noise model? The key point here is that we are contrasting our model with the observations. This is very different from the case of indistinguishable states, where we are contrasting various model trajectories with each other and asking whether or not we are likely to be able to distinguish one from another given only noisy observations. In this case, one considers all possible realisations of the observational noise. Thus when working with indistinguishable states we consider model trajectories and the statistics of the observational noise, whereas with shadowing we contrast a model trajectory and the actual set of observations in hand. Shadowing has a long history in non-linear dynamical systems dating back to the early 1960s; a discussion of the various casts of shadow can be found in Smith (2001).

In terms of actually constructing an operational ensemble, the relevant members of the indistinguishable sets are those that have, in fact, shadowed given the particular realisation of the observational noise in the recent past. More clearly: recall that each set of indistinguishable states is determined by integrating over all possible realisations of the observational noise; even when we wish to pick our ensemble members from this set, we will weight them with respect to the observations obtained in the one realisation of the noise which we have access to (our observations).

Given the arguments above, it follows that within the perfect model scenario an infinite number of distinct, infinitely long shadowing model trajectories would exist, each trajectory shadowing observations from the beginning of time up until the present. These trajectories are easily distinguished from each other within the model state space, but the noisy observations do not contain enough information to identify which one was used to generate the data. The contents of this set of indistinguishable states will depend on the particular realisation of the observational noise, but the set will always include the generating trajectory (also known as 'Truth'). This fact implies that even if granted all the powers of our twenty-first century demon, we would still have to make probabilistic forecasts. Epistemologically, one could argue that the 'true state' of the system is simply not defined at this point in time, and that the future is no more than a probability distribution. Accepting this argument implies that after each new observation is made, the future really ain't what it used to be.

Of course, even this restriction to an ever-changing probabilistic future is a difficulty to which we can only aspire. We do not have a perfect model; we do not even have a perfect model class, meaning that there is no combination of model parameters (or available parametrisations) for which any model accessible to us will provide a shadowing trajectory. Outside the perfect model scenario (PMS), the set of indistinguishable states tends to the empty set (Judd and Smith, 2004). In other words, the probability that the data was generated from *any* model in the model class approaches zero as we collect more observations, an awkward fact for the applied Bayesian.[4] But accepting this fact allows us to stop aiming for the perfect model, just as accepting uncertainty in the initial condition freed us from the unattainable goal of a single accurate forecast trajectory from 'inaccurate' starting conditions. When the model is very good (that is, the typical i-shadowing times are long compared with the forecast time) then the consideration of indistinguishable states suggests a new approach to several old questions.

9.5 Indistinguishable states and the fair valuation of forecasts

In this section we will consider the aims of model building and the evaluation of ensemble forecasts. Rather than repeat arguments in Smith (1997, 2001) on the importance of distinguishing model-variables from observed variables, we will consider the related question of the 'best' parameter value for a given model class. Arguably there is no such thing outside of PMS, and the insistence on finding a best can degrade forecast performance. We will then consider the game of weather roulette, and its use as an illustration for the economic decisions Charles and Charlotte make every day. A fair comparison of the value of different forecasts requires contrasting like with like, for example we must 'dress' both BFG and EPS simulations in order to obtain a fair evaluation of the probabilistic forecasts available from each method. Weather roulette also allows us to illustrate Charles' favoured cost function for economic forecasts, the logarithmic skill score called *ignorance* (Roulston and Smith, 2002). Relative ignorance can also be used to obtain insight on operational questions such as the division of computational resource between ensemble size and model resolution for a given target, as illustrated with Heathrow temperatures below (see also Smith *et al.*, 2001). Two worked economic examples are discussed in Section 9.7.

9.5.1 Against best

What parameter values should best be used in a forecast model? If the system that is generating the data corresponds to a particular set of parameters in our forecast

model, then we have a *perfect model class*; obviously that set of parameters would be a candidate for 'best'. Outside PMS, at least, the question of best cannot be decoupled from the question of why we are building the model. We will duck both questions, and simply consider a single physically well-understood parameter, the freezing point of water,[5] within three (imperfect) modelling scenarios.

At standard pressure, it is widely accepted that liquid water freezes at zero degrees C. In a weather model with one millimetre resolution, I would be quick to assign this value to the model-freezing point. In a weather model with one-angstrom resolution, I would hope the value of zero degrees C would 'emerge' from the model all by itself. And at 40 kilometre resolution? Well at 40 km resolution I have no real clue as to the relation between model-temperature and temperature. I see no defensible argument for setting this parameter to anything other than that value which yields the best distribution of shadowing trajectories (that is, the distribution which, in some sense, reflects the longest shadowing times; a definition we will avoid here).

Of course, the physical relevance of the mathematical form used for the model parametrisation assumes that the parameter value lies in some range; internal consistency (and relevance) of the model parametrisation itself places some constraints on the values of the model parameters within it. This suggests, for example, that the freezing point of water should be somewhere around zero, but does not suggest any precise value.

This confusion between model-parameters and their physical analogues, or even better between model-variables and the observations (direct empirical measurements), is common. The situation is not helped by the fact that both are often given the same name; to clarify this we will distinguish between temperature and model-temperature where needs be.

Translating between model variables and observables is also related to representation error. Here we simply note that representation error is a shortcoming of the model (and the model-variables) not the observations. A reanalysis is a useful tool, but its value derives from the observations. In fifty years' time the temperature observations recorded today at Heathrow airport will still be important bits, whereas no one will care about model-temperature at today's effective grid resolution. The *data*, not the model-state space, endure.

Outside PMS, it is not clear how to relate model-variables to variables. To be fair, one should allow each model its own projection operator. Discussion of the difficulties this introduces will be pursued elsewhere, but there is no reason to assume that this operator should be one-to-one; it is likely to be one-to-many, as one model state almost surely corresponds to many atmospheric states if it corresponds to any. It may even be the case that there are many atmospheric states for each model state, and that each atmospheric state corresponds to more than one model state: a many-to-many map. But again, note that it might be best to avoid the assumption that there is 'an' atmospheric state altogether. We do not require this assumption. All we ever have are model states and integer observations.

9.5.2 Model inadequacy: the biggest picture

Let us now reconsider the issues of forecasting within the big picture. Once again, suppose for the moment that there exists some True state of the atmosphere, call it $\mathbf{X}(t)$. This is the kind of thing Laplace's nineteenth-century demon would need to know to make perfect forecasts, and given $\mathbf{X}(t)$ perfect BFG forecasts would follow even if the system were chaotic. As Laplace noted, mere mortals can never know $\mathbf{X}$, rather we make direct measurements (that is, observations) s, which correspond to some projection of $\mathbf{X}$ into the space of our observing system. Given s, and most likely employing our model $\mathbf{M}$ as well, we project all the observations we have into our model-state space to find a distribution for $x(t)$. Traditionally attention has focused on one model state, often called the analysis, along with some uncertainty information regarding the accuracy of this state. This is nothing more than data assimilation, and the empirically relevant aim of data assimilation is an ensemble (or probability distribution function), not any single state. Using our model, we then iterate $x(t)$ forward in time to some future state, where we reverse the process (via downscaling or model output statistics) to extract some observable w. In the meantime, the atmosphere has evolved itself to a new state, and we compare our forecast of w with our new observation and the corresponding observed projection from the new $\mathbf{X}$.

Although pictures similar to the one drawn in the preceding paragraph are commonplace, the existence of Truth, that is the existence of $\mathbf{X}$, is mere hypothesis. We have no need of that hypothesis, regardless of how beautiful it is. All we ever have access to are the observations, which are mere numbers. The existence of some 'True' atmospheric state, much less some mathematically perfect model under which it evolves, is untestable. While such questions of belief are no doubt of interest to philosophers and psychologists, how might they play a central role within science itself? In questions of resource allocation and problem selection, they do.

Accepting that there is no perfect model is a liberating process; perhaps more importantly, it might allow better forecasts of the real atmosphere. Doing so can impact our goals in model improvement, for example, by suggesting the use of ensembles of models and improving their utility, as opposed to exploring mathematical approximations which will prove valid only for some even better model which, supposing it exists, is currently well beyond our means. The question, of course, is where to draw the line; few would question that there is significant information content in the 'Laws of Physics' however they lie, yet we have no reliable guide for quantifying this information content. In which cases will PMS prove to be a productive assumption? And when would we do better by maintaining two (or seven) distinct but plausible model structures and *not* trying to decide which one was best? What is the aim of data assimilation in the multimodel context? Indeed, is there a single aim of data assimilation in any context: might the aim for nowcasting differ from that of medium-range forecasting?

Charlie, Charles and Charlotte, along with others in the economy rather than in economics, seem untroubled by many of these issues; they tend to agree on the same cost function, even if they agree on nothing else. So we will leave these philosophical issues for the time being, and turn to the question of making money.

9.5.3 Weather roulette

A major goal of this chapter is to convince you that weather roulette is not only a reasonable illustration of the real-world economic decisions that Charles and Charlotte deal with on a daily basis, but that it also suggests a relevant skill score for probabilistic forecast systems. Weather roulette (M. S. Roulston and L. A. Smith, unpublished data) is a game in which you bet a fixed stake (or perhaps your entire net worth) on, say, the temperature at Heathrow. This is repeated every day. You can place a wager on each number between -5 and 29, where 29 will include any temperature greater than 29 and -5 any value below -5. How should you spread your bet? First off, we can see that it would be foolish not to put something on each and every possibility, just to be sure that we never go bust. The details of the distribution depend on your attitudes toward risk and luck, among other things; we will consider only the first. In fact we will initially take a risk neutral approach where we believe in our forecast probabilities: in this case we distribute our stake proportionally according to the predicted probability of each outcome. We imagine ourselves playing against a house that sets probabilistically fair odds using a probability distribution different from ours. Using this approach we can test the performance of different probability forecast systems in terms of how they fare (either as house or as punter).

As a first step, let's contrast how the ECMWF ensemble forecast would perform betting against climatological probabilities. Given a sample-climatology based on 17 years of observations, we know the relative frequency with which each option has been observed (given many centuries of data, we might know this distribution for each day of the year). But how do we convert an ensemble forecast of about 50 simulations into a probability forecast for the official observed temperature at Heathrow airport? There are a number of issues here.

9.5.4 Comparing like with like

How can Charlotte contrast the value of two different probability forecasting systems, say one derived from a high-resolution BFG forecast and the other from an EPS forecast? Or perhaps two EPS forecasts which differ either in the ensemble formation method or as to the ensemble size? There are two distinct issues here: how to turn simulations into forecasts and how to agree on the verification.

The first question is how to turn a set of model(s) simulations into a probability weather forecast. In the case of ensembles there are at least three options: fitting some parametric distribution to the entire ensemble, or dressing the individual ensemble's members (each with an appropriate kernel), or treating the entire forecast ensemble

(and the corresponding verification) as a single point in a higher dimensional space and then basing a forecast upon analogues in this product space. We will focus on dressing the ensemble, which treats each individual member as a possible scenario. The product space approach treats the statistics of the ensemble as a whole; examples in the context of precipitation and wind energy can be found in Roulston *et al.* (2001, 2003).

Treating the singleton BFG as a point prediction (that is, a delta function) does not properly reflect its value in practice. To obtain a fair comparison with an EPS, we 'dress' the forecast values by replacing each value with a distribution. For the singleton BFG, one can construct a useful *kernel*[6] simply by sampling the distribution of historical 'error' statistics. The larger the archive of past forecast-verification pairs is, the more specific the match between simulation and kernel. Obviously we expect a different kernel for each forecast lead-time, but seasonal information can also be included if the archive span is long enough. Dressing the BFG in this way allows a fair comparison with probability forecasts obtained from ensemble forecasts. If nothing else, doing so places the same shortcoming due to counting statistics on each case. To see the unfairness in treating a BFG forecast as if it were an ensemble which had forecast the same value 50 times, recall the game of roulette. Think of the advantage one would have in playing roulette if, given an accountable probability forecast, you could place separate bets with 50 one dollar chips on each game, rather than having to place one 50 dollar chip on each game. To make a fair comparison, it is crucial that the single high-resolution simulation is dressed and not treated as an unequivocal forecast.

Suppose we have in hand a projection operator that takes a single model state into the physical quantity we are interested in, in this case the temperature at Heathrow. Each ensemble member could be translated into a specific forecast; because the ensemble has a fixed number of members, we will want to dress these delta functions to form a probability distribution that accounts both for our finite sample and for the impact of model error. We can use a best member approach to dressing, which is discussed in Roulston and Smith (2003) or simply use Gaussian kernels. Why don't we dress the ensemble members with the same kernel used for the BFG? Because to do so would double count for uncertainty in the initial condition. The only accounting for this uncertainty in the BFG is the kernel itself, while the distribution of the ensemble members aims to account for some of the uncertainty in the initial condition explicitly. Thus, other things being equal, the BFG kernel is too wide for ensemble members. Which kernel is appropriate for the ensemble members? Just as the error in the high-resolution forecast is too wide, the error in the ensemble mean is irrelevant. By using a kernel based on the best member of the ensemble, we aim to use a kernel that is as narrow as possible, but not more so. We do not, of course, know which member of the ensemble will turn out to provide the best forecast, but we do know that one of them will. Of course one must take some care in identifying the best member: considering the distant future of a trajectory will obliterate the more relevant information in the recent past. The critical issue here is to dress the simulations, the ideal kernel may well depend upon the circumstances, the size of the forecast archive, for instance.

The necessity of dressing emphasises the role of the forecast archive in the proper valuation validation of EPS forecasts. It also shows that the market value of an EPS will be diminished if a suitable EPS forecast archive is not maintained. This remains the case when we use simple kernels, such as a Gaussian distribution, and use the archive to determine any free parameters.

The second question of model verification addresses how we come to agree on what actually happened. This is not as clear-cut as it might first appear; a cyclist may arrive home drenched only to learn that, officially, it had not rained. Often, as in the case of the cyclist, what 'happened' is decided by decree: some agency is given the power to define what officially happened. Charles is happy with this scenario as long as the process is relatively fair: he wants to close his book at the end of each day and go home. Charlotte may be less pleased if the 'official' forecast says her generators were running at 77%, while a direct calculation based on the amount of gas burned yields 60%. In part, this difference arises because Charles really is playing a game, while Charlotte is trying to do something real. Outside the perfect model scenario, there appears to be no unique optimal solution.

9.5.5 Ignorance

Roulston and Smith (2002) discuss an information theoretical skill score that reflects expected performance at weather roulette. Called *ignorance*,[7] this skill score reflects the expected wealth doubling time of two gamblers betting against each other using probability forecasts, each employing Kelly systems.[8] For instance, a difference in ignorance scores of one bit per game would suggest that, on average, the players with lower ignorance score would double their stake on each iteration. The typical wealth doubling (or halving) time of a balanced roulette table is about 25 games, favouring the house, while scientific roulette players have claimed a doubling time on the order of three and favouring the player. How do these values arise?

Prior to the release of the ball, the probability of each number on a balanced roulette wheel can be taken to be equal, that is 1 in 37 in Europe (and 1 in 38 in the United States). If the house offers odds of 36-for-1, then the expected value of a unit stake after a single bet on a single number is 36/37. Raising this to the 25th power yields the reduction of the initial stake to one half; hence the expected wealth halving time is, in this case, roughly 25 games.

Roulette is a particularly appropriate analogy in that bets can be placed after the ball has been released. If, using observations taken after the ball is released, one could predict the quarter of the wheel on which the ball would land, and if such predictions were correct one third of the time, then one would expect a pay-off of roughly 36/9 once in every three games. This is comparable to reports of empirical returns averaging 30% or the expected stake doubling time of about three noted above.

There are several important things to note here. First, the odds offered by the house are not fair, even if the wheel is. Specifically, the implied 'probabilities' (the reciprocal of the odds) over the available options do not sum up to one: $37 \times (1/36) > 1$.

In practice it is rarely, if ever, the case that this sum is one, although this assumption is built into the most common definition of fair odds.[9]

Second, ignorance contrasts probability forecasts given only a single realisation as the verification. That is, of course, the physically relevant situation. In the 1995 ECMWF Seminar, we discussed the difference between the forecast distribution and some 'true' distribution as quantified via the Kolmogorov–Smirnov (KS) distance; a more elegant measure of the difference between two probability distributions is provided by the *relative entropy* (see Kleeman, 2002). What these and similar approaches fail to grasp, however, is that there is no 'true' distribution to contrast our forecast distribution with. Outside PMS, our forecast distribution rarely even asymptotes to the climatology unless forced to (and if forced, it asymptotes to a sample-climatology).

In the case of roulette, each spin of the wheel yields a single number. Before the ball is released, a uniform prior is arguably fair; but the relevant distribution is defined at the point when all bets must be placed, and this is *not* uniform. And there's the rub. What is a fair distribution at this point? Even if we assume it exists, it would depend on the size of your computer, and your knowledge of mechanics, and the details of this particular wheel, and this particular ball. And so on. Computing the relative entropy (or the KS distance) requires knowledge of both the forecast distribution and the true distribution conditioned on the observations. The equation that defines the relative entropy has two terms. The first term reflects the relative frequency with which certain forecast probabilities are verified; the second term reflects the relative frequency with which certain probabilities are forecast by a perfect model. Thus outside PMS the second term is unknowable, and hence the relative entropy is unavailable. The first term reflects the ignorance skill score. A shortcoming of ignorance is that it provides only a relative skill score ranking alternatives. Its advantage, of course, is that it is deployable in practice. As we have only the data, and data are but numbers, only limited skill scores like ignorance are available in practice. Similarly, our real-time performance will depend on the single future we experience, not the expectation over some many-worlds collection of 'possibilities'.

9.5.6 Heathrow temperature

So back to weather roulette: How does the dressed ensemble fare against climatology? Rather well. Consider daily bets, each with a stake of £100, placed over the period of a year starting on 23 December, 1999. The three-day forecast based on the ECMWF ensemble made almost £5000 in this period, while the ten-day forecast made about £1000 over the same period. As expected even in a perfect model, the value of the ensemble relative to climatology decreases with lead time. And there is more. The dressed ECMWF ensemble can also be used to bet against house odds based on the (dressed) ECMWF high-resolution best first guess (BFG) forecast. In this case the relative information gain is greater at a lead time of 10 days than at three, with winnings of over £5000 and over £2000, respectively. This also makes sense,

inasmuch as the value of the ensemble forecast relative to climatology is greatest in the short range, while its value relative to a high-resolution forecast is greater at longer lead times (M. S. Roulston and L. A. Smith, unpublished data).

But there is still more, we can contrast the 51 member ensemble with the 10 member ensemble (all ensembles are dressed before use; the best-member kernel will, of course, vary with the size of the ensemble); is there statistically significant value added in the 51 member case relative to the case of randomly selecting 10 members? Yes. Relative to the 12-member case? Yes. And the 17-member case? No. Arguably, if Charles were betting only on Heathrow temperatures he would have done as well dressing 17 members as with using (that is, buying) all of them. He would have taken them for free (why not?) but not paid much for them. There is nothing universal about the number 17, in this context. The relevant ensemble size will depend on the details of the target variable as well as the model.

Of course any meteorological sales person worth their salt would immediately point out that users like Charlotte are interested in conditional probabilities of multiple variables at multiple locations. Charlotte is interested in the efficiency of each of her CCGT plants, as well as some integrated measure of electricity demand. But Charles is happy to stick with Heathrow, as long as his probability forecasts are making him money and costing him as little as possible. Making/pricing complicated multi-site derivatives is hard work; and he knows that it is dangerous as well: the curse of dimensionality implies tight constraints on the conditional probability forecasts that can be pulled out of even the best Monte Carlo ensemble with only a few dozen, or a few thousand, members. Charles simply need not take these risks, if the market in trading Heathrow temperatures is both sufficiently profitable and sufficiently liquid. Charlotte, by contrast, is exposed to these risks; the best she can do is understand the limits placed on the available forecasts by current technology.

There are three important take-home messages here: first that, like it or not, the ideal distribution between ensemble size and resolution will be target (that is, user) dependent; second that the marginal value of the $n + 1$st ensemble member will go to zero at different times for different uses; and third that the value of the EPS as a whole will depend critically upon the provision of an archive that allows users to convert these simulations into forecasts of the empirical quantities of interest.

9.6 Lies, damn lies, and the perfect model scenario

> *For the materialist, science is what teaches us what to believe. For the empiricist, science is more nearly what teaches us how to give up our beliefs.*
>
> Bas van Fraassen

Philosophers might call the meteorologist striving for PMS a realist; he believes in the physical reality of the model states and in the truth, or approximate truth, of

the model. Alternatively, one modern variety of empiricist aims only for empirical adequacy. Such labels might seem inappropriate outside a philosophy department, if they were not relevant for the allocation of resources and hence the progress of science.

A model is empirically adequate if the dynamics of the model are in agreement with the observations. In this context, the word prediction often refers to prophecies as well as forecasts (see Smith, 1997). To judge empirical adequacy requires some accounting for observational noise and the fair use of a projection operator between the model-variables and the observables; *i*-shadowing would be a necessary condition for empirical adequacy of a dynamical system. There is still much to be understood in this direction, nevertheless it is not clear to me that *any* of our dynamical models are empirically adequate when applied to dynamic (non-transient) physical systems. There are a number of points, however, where the decision to work within PMS impacts the relevance of the statistics used to judge our models and the choice of how to 'improve' them.

9.6.1 Addressing model inadequacy and multiple model ensembles

The philosophical foundations of theories for objective probability distributions are built about the notion of equally likely cases or events (see Gillies, 2000). Within PMS, the perfect ensemble is an invocation of this ideal for chaotic dynamical systems that are perfectly modelled but imperfectly observed. This view of the world is available only to our twenty-first century demon who has access to various sets of indistinguishable states. Given a collection of good but imperfect models, we might try and use them simultaneously to address the issue of model inadequacy. But once we employ multiple imperfect models, the epistemological foundations that justify empirical probability forecasting turn to sand. While we can tolerate uncertainty in the parameter values of a model that comes from the correct model class by invoking what are effectively Bayesian methods, it appears we cannot find an internally consistent framework to support *objective* (empirically relevant) probability forecasts[10] when using multiple models drawn from distinct model classes, the union of which is known to be imperfect. Of course one can draw comfort in those aspects of a forecast in which models from each model class independently assign similar probabilities; but only comfort, not confidence (see Smith, 2002). Both systematic and flow dependent differences in the skill scores between the probabilistic forecasts from each model class may help us identify which physical phenomena deserve the most attention in each model class. Thus with time we can improve each of the models in the ensemble, while our probabilistic forecasts remain infinitely remote from the accountable probability forecasts of our twenty-first century demon.

9.6.2 Model inadequacy and stochastic parametrisations

Every attempt at model improvement is an attempt to reduce either parameter uncertainty or model inadequacy. *Model inadequacy* reflects the fact that there is no model within the class of models available to us that will remain consistent with all the data. For example, there may be no set of parameter values that enable a current model to shadow the observations *out-of-sample*[11] (or even in-sample?). Finding the (metric dependent) 'best' parameters, or distributions of parameters, and improving the data assimilation scheme are attempts at minimising the effects of model inadequacy within a given class of models. But model inadequacy is that which remains even when the best model within the model class is in hand; it affects both stochastic models and deterministic models.

Historically, physicists have tended to employ deterministic models, and operational numerical weather prediction models have been no exception to this trend. There are at least two good reasons why our forecast models should be stochastic in theory. The first comes from recent results (Judd and Smith, 2004) which establish that, given an imperfect non-linear chaotic model of a deterministic system, better state estimation (and perhaps, even better probabilistic forecasts) can almost certainly be obtained by using a stochastic model even when the system which generated the data really is deterministic! The second is the persuasive argument that, given current model resolution, it makes much more sense (physically) to employ stochastic subgrid-scale parametrisations than to employ dogmatically some mean value, even a good estimate of the expected value (Palmer, 2001; Smith, 2002). And in addition to these theoretical arguments, stochastic parametrisations have been shown to be better in practice (Buizza *et al.*, 1999). It is useful to separate arguments for improving a forecast based on each of these two reasons; we should maintain the distinction between methods which improve our model class (say, by adding stochastic physics) and those that deal with residual model inadequacy (which will always be with us).

While adopting stochastic parametrisations will make our models fundamentally stochastic, it neither removes the issue of model inadequacy nor makes our model class perfect. Consider what is perhaps the simplest stochastic model for a time series: independent and identically distributed (IID) normal (Gaussian) random variables of mean zero and standard deviation one. Data are numbers. *Any* data set has a finite (non-zero) probability of coming from this trivial IID model. Adjusting the mean and the standard deviation of the model to equal the observed sample-mean and sample-deviation will make it more difficult to reject the null hypothesis that our IID 'model' generated the data, but not much more difficult. We are soon faced with probabilities so low that, following Borel (1950), we could say *with certainty* that even a stochastic model does not shadow the observations. The possibility to resemble differs from the ability to shadow.

One often overlooked point here is that whenever we introduce a stochastic element into our models we also introduce an additional constraint: namely that to be a shadowing trajectory, the innovations must be consistent with the stochastic process specified in the model. Stochastic parametrisations may prove a tremendous improvement, but they need not yield *i*-shadowing trajectories even if there exist *some* series of innovations that would produce trajectories similar to a time series of analysis states. To be said to shadow, the particular series of innovations must have a reasonable probability given the stochastic process. Experience suggests that it makes little difference if we require a 95% isopleth, or 99%, or 99.999% for that matter. Model inadequacy manifests itself rather robustly. Without this additional constraint the application of the concept of *i*-shadowing to stochastic models would be trivial; the concept would be useless as any rescaled IID process could be said to shadow any set of observations. With this constraint, the introduction of stochastic terms does not guarantee shadowing, and their contribution to improved probabilistic forecasts can be fairly judged.

Even within PMS we must be careful that any critical, theoretically sound assumptions hold if they are relevant in the particular case in question (Hansen and Smith, 2000; Gilmour *et al.* 2001). Outside PMS a model's inability to shadow holds implications for operational forecasting, and for the Bayesian paradigm in applied science if not in mathematics. First of all, it is demonstrable that we can work profitably with imperfect models full of theory-laden (or better still, *model-laden*[12]) variables, but we can also be badly misled to misallocate resource in the pursuit of interesting mathematics, which assumes an unjustified level of perfection in our models. Ultimately, only observations can adjudicate this argument – regardless of what we 'know' must be the case.

9.6.3 Dressing ensemble forecasts and the so-called 'superensemble'

The superensemble method introduced by Krishnamurti *et al.* (1999) is a very interesting method for extracting a single 'locally optimised' BFG forecast from a multimodel ensemble forecast. In short, one finds the optimal weights (in space, time, and target variable) for recombining an ensemble of multimodel forecasts so as to optimise some root-mean-square skill score of the resulting BFG forecast.

The localised relative skill statistics generated within this 'superensemble' approach must contain a wealth of data of value in understanding the shortcomings of each component model and in addressing these model inadequacies. Nevertheless the 'superensemble' approach aspires only to form a single BFG forecast, and thus it might be more aptly called a 'super ensemble-mean' approach. How might we recast the single forecast output from the 'superensemble' approach, in order to make a 'like with like' comparison between the 'superensemble' output and a probability

forecast? The obvious approach would be to dress it, either with some parametric distribution or in the same way we dressed the high-resolution forecasts in Section 9.5 above. Hopefully the shortcoming of either approach is clear: by first forming a single super ensemble-mean, we have discarded any information in the original distribution of the individual model states. Alternatively, dressing the individual ensemble members retains information from their distribution. So, while it is difficult to see how any 'superensemble' approach could outperform either a dressing approach or a product space method (or any other method which retains the distribution information explicitly), it would be interesting to see if in fact there is any relevant information in this distribution!

9.6.4 Predicting the relevance of indistinguishable states

The indistinguishable states approach suggests interesting alternatives both to current methods of ensemble formation and to the optimised selection of additional observations (Judd and Smith, 2001, 2004). The second are often called *adaptive observations* since the additional observation that is suggested will vary with the current state of the atmosphere (see Lorenz and Emanuel, 1998; Hansen and Smith, 2000).

Ensemble formation via indistinguishable states avoids the problems of adding finite perturbations to the current best guess analysis. The idea is to direct computational resources towards maintaining a very large ensemble. Rather than discarding ensemble members from the last initialisation some would simply be reweighted as more observations are obtained (see also Beven, 2002). It would relax (that is, discard) the assumptions of linearised uncertainty growth, for example that the observational uncertainty was small relative to the length scale on which the linearisation is relevant (see Hansen and Smith, 2000; Judd, 2003), or that the uncertainty distributions are Gaussian. And, by making perturbations as far into the distant past as possible, the ensemble members are as consistent with the long-term dynamics as possible; there are no unphysical 'balance' issues. Perhaps most importantly, an indistinguishable states approach appears to generalise beyond the assumption that near shadowing trajectories of reasonable duration do, in fact, exist when it is difficult to see how any of the current alternative approaches might function in that case. Much work remains to be done in terms of quantifying state dependent systematic model error (such as drift, discussed by Orrell *et al.* 2001) and detecting systematic differences between the behaviour of the analysis and that of the ensemble and its members.

When required, the ensemble would be reseeded from additional trajectories initiated as far back in time as practical, unrealistic perturbations would be identified and discarded without ever being included in a forecast. Reweighting evolved ensemble members given additional data (rather than discarding them), allows larger ensembles to be maintained, and becomes more attractive both as the period between initialising weather ensembles decreases and in the case of seasonal ensembles where many observations may be collected between forecast launches. In the former case at least,

we can form lagged ensemble ensemble forecasts (LEEPS) by reweighting (and perhaps changing the kernel of) older ensemble members if they either remain relevant in light of the current observations or contribute to the probability forecast at any lead-time. Of course, outside PMS the relative weighting and the particular kernel assigned to the older members can differ from that of the younger members, and it would be interesting to use the time at which this weighting went to zero in estimating an upper bound for a reasonable maximum forecast lead time. And outside PMS, one must clearly distinguish between the ensemble of model simulations and a probabilistic forecast of weather observables.

Seasonal forecasts as studied within DEMETER (see www.ecmwf.int/demeter) provide ensembles over initial conditions and model structure. While it may prove difficult to argue for maintaining multiple models within PMS, the need to at least sample *some* structural uncertainties outside PMS provides an *a-priori* justification for multimodel ensembles, as long as the various models are each plausible. Indeed, it is the use of a single model structure that can only be justified empirically in this case, presumably on the grounds that, given the available alternatives, one model structure is both significantly and consistently better than all the others.

Within PMS, ensembles of indistinguishable states based on shadowing trajectories aims to yield nearly accountable probability forecasts, while operational methods based on singular vector or on bred vector perturbations do not have this aim, even in theory. The indistinguishable states framework also suggests a more flexible approach to adaptive observations if one model simulation (or a set of simulations) was seen to be of particular interest. To identify adaptive observations one can simply divide the current trajectories into two groups, one group in which each member has the interesting property (for example, a major storm) and the other group in which the simulations do not; call these groups red and blue. One could then identify which observations (in space, time, and model-variable) are most likely to provide information on distinguishing the distribution of red trajectories from the distribution of blue, or better said: which observations are most likely to give members of one of the groups high probability and those of the other low probability. As additional regular observations are obtained, the main computational overhead in updating our method is to reweight the existing trajectories, a relatively low computational cost and an advantage with respect to alternative approaches (see Hansen and Smith, 2001, and references thereof). Given a multimodel multi-initial condition ensemble, the question of adaptive observations shifts from complex assumptions about the growth of uncertainty under imperfect models, to a question of how to best distinguish between two subsets of known trajectories.

With multimodel forecasts we can also use the indistinguishable states framework to select observations that are most likely to 'falsify' the ensembles from one of two models on a given day. To paraphrase John Wheeler: each of our models is false; the trick is to falsify them as quickly as possible. Here only the observations can adjudicate; while what we decide to measure is constrained by what we can think

of measuring, the measurement obtained is not. Le Verrier was as confident of his prediction of Vulcan as he was of his prediction of Neptune, and while both planets were observed for some time only Neptune is with us today.

Just as the plural of datum is not information, the plural of good idea is not theoretical framework. The indistinguishable states approach to forecasting and predictability has significant strengths over competing strategies, but its operational relevance faces a number of hurdles that have yet to be cleared. This statement should not be taken to indicate that the competition has cleared them cleanly!

9.6.5 A short digression toward longer timescales: the in-sample science

Noting the detailed discussion in the chapters by Tim Palmer and Myles Allen, we will not resist a short digression towards longer timescales (additional discussion can be found in Allen, 1999, Smith 2002, Stainforth *et al.*, 2005 and the references therein). An operational weather model makes many 7-day forecasts over its short lifetime; contrast this case with that of a climate model used to make 50-year forecasts, yet considered obsolete within a year or two. The continuous feedback from making forecasts on new unseen (out-of-sample) data is largely denied the climate modeller, who is constrained by the nature of the problem forever to violate one of the first maxims of undergraduate statistics: never restrict your analysis to in-sample statistics. By construction, climate modelling is an in-sample science. And the fundamentally transient nature of the problem makes it harder still.

As argued elsewhere (Smith, 2002), an in-sample science requires a different kind of consistency constraint. If model inadequacy foils our attempts at objective probability forecasts within the weather scenario, there is little if any chance of recovering these in the climate scenario.[13] We can, however, interpret multiple models in a different way. While approaches like best-member dressing can take into account the fact that different models will perform better in different conditions in the weather context, a climate modeller cannot exploit such observations. In the weather scenario we can use all information available at the time the models are launched when interpreting the distribution of model simulations, and as we launch (at least) once a day, we can learn from our mistakes.

Inasmuch as climate forecasting is a transient experiment, we launch only once. It is not clear how one might combine a collection of single runs of different climate models into a sensible probability forecast. But by studying the in-sample behaviour of ensembles under a variety of models, we can tune each model until an ensemble of initial conditions under each and every individual model can at least bound the in-sample observations, say from 1950 to 2000. If ensembles are then run into the future, we can look for the variables (space and time scales) on which the individual model ensembles bound the past and agree (in distribution) regarding the future.

Of course agreement does not ensure relevance to the future; our collection of models can share a common flaw. But if differences in the subtle details of different models that have similar (plausible) in-sample performance are shown to yield significantly different forecast distributions, then no coherent picture emerges from the overall ensemble. Upon adding a new model to this mix, we should not be surprised by a major change to the overall forecast distribution. This may still occur even if 'each and every one' of the current models has similar forecast distributions, but in this case removing (and hopefully, adding) one model is less likely to do so. In any event, we can still use these differences in the forecast distributions to gain physical insight, and improve each model (individually) using the in-sample data yet again (Smith, 2002). But as long as the details can be shown to matter significantly, we can form no coherent picture.

9.7 Socio-economic relevance: why forecast everyday?

From a scientific point of view, it is interesting to ask why we make forecasts every day? Why not spend all available funds making detailed observations in, say, January, and then spend the rest of the year analysing them, using the same computational resources but with a higher resolution model than possible operationally? Once we got the physics of the processes for January down pat, we could move on to February. And so on. There are a number of reactions to this question, but the most relevant here is the simple fact that numerical weather forecasting is more than just a scientific enterprise; real-time forecasting is largely motivated by the socioeconomic benefits it provides. One of the changes we will see in this century is an increase in the direct involvement of users, both in the consideration of their desires and in the exploitation of their data sets. Closing this loop can benefit both groups: users will employ forecast products more profitably while modellers will have to leave the 500 mb model-pressure height field behind, along with the entire model state space, and again give more consideration to empirically accessible variables.

Electricity demand provides real-time observations reflecting a number of environmental variables, updated on the timescale of a model time step, and spatially integrated over areas similar to model grid spacing. Might not assimilating this data (or using it to reweight the trajectories of what were indistinguishable) be more likely to yield relevant information than a single thermometer accurate to 16 bits? Charlotte uses demand observations every day; she would certainly be willing to sell (or barter) these bits for cheaper (or better) forecasts.

It is easily observed that many talented meteorologists dismiss the idea of a two-way exchange with socio-economics out of hand. They state, rather bluntly, that meteorologists should stick to the 'real science' of modelling the atmosphere, even dismissing the comparison of forecast values with observations as mere 'post-processing'.[14] Interestingly, if only coincidentally, these same scientists are

often those most deeply embedded within PMS. Of course I do not really care about forecasting today's electricity demand *per se* any more than I am interested in benthic foraminifera; but I do care very much about empirical adequacy and our ability to forecast things we can actually measure: any empirically accessible quantity whatever its origin. I am not overly concerned whether a quantity is called 'heating degree days' or 'temperature'. As long as they correspond to something we can measure, not a model variable, I'll take as many as I can get. Probabilistic forecasts for both temperature and heating degree days, as well as cumulative heating degree days, are posted daily on www.dime.lse.ac.uk.

9.7.1 Ensembles and wind power

So let us consider three examples of economic meteorology. The first is a study of a hypothetical wind farm using real wind data, real forecasts, real electricity prices and a non-existent wind farm just south of Oxford. (Detailed results are available in Roulston *et al.*, 2003.) The economic constraints vary with changes in regulation, which occur almost as frequently as changes in an operational weather model. In our study, the wind farm must contract for the amount of electricity it will produce in a given half hour a few days in advance; it will not be paid for any overproduction, and will have to supply any shortfall by buying in electricity at the spot price on the day. What we can do in this example is to contrast several different schemes for setting the size of the contract, and then evaluate them in terms of the income of our fictional wind farm. Contrasting the use of climatology, the dressed ECMWF high-resolution forecast and the dressed ECMWF ensemble forecast shows, for instance, that at day 4 there is a clear advantage in using the ensemble. Would Charlotte buy the ECMWF ensemble? That question involves the cost of the ensemble relative to its benefits, the cost of the high-resolution run, the size of the current forecast archive and the availability of alternative, less expensive probability forecasts. But it is the framework that is important here: she can now make an economic choice between probability forecasts, even if some of those probability forecasts are based on single BFG model runs (singleton ensembles). As shown in our next example, it is likely that in some cases the dressed ensemble forecast may not contain significantly more information than the dressed high-resolution forecast. Still, the framework allows us to see when this is the case, and often why. Figure 9.1 shows the daily income from the wind farm: the upper panel uses climatology forecasts; the lower panel uses the dressed ensemble (for more detail, see Roulston *et al.*, 2003). The ensemble forecast provides increased profit at times of unseasonable strong winds (March 2000) while avoiding loss at times of unseasonably weak winds (January 2000). It is not perfect, of course, just better. Presenting the impact of weather forecasts in this format allows Charlotte to make her own decisions based on her company's attitude to risk.

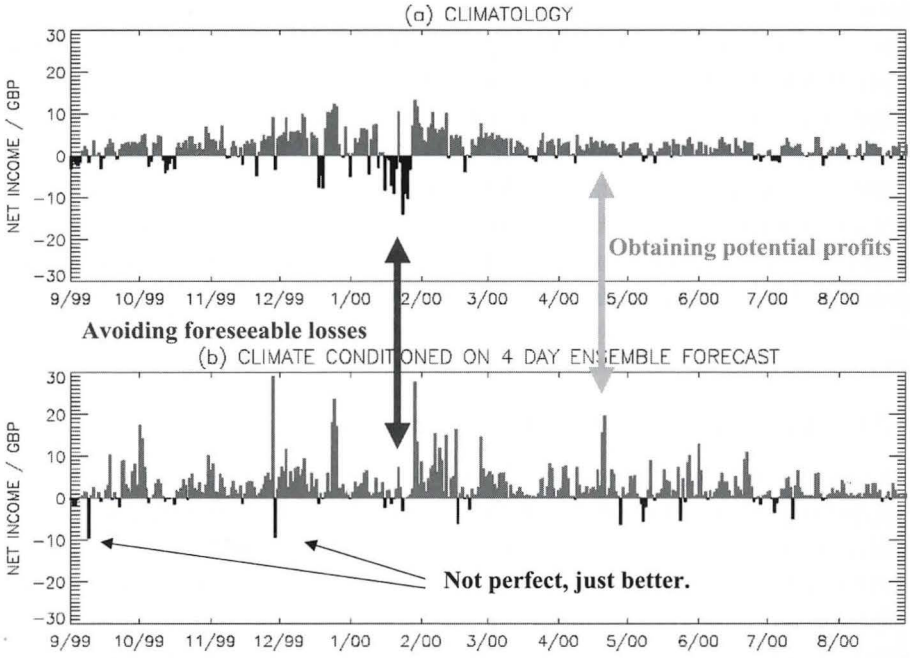

Figure 9.1 The daily income from a hypothetical wind farm based on observed winds, real electricity prices and ECMWF forecasts. (a) The profit when the estimated production use is based on climatology. (b) The same when based on the ECMWF ensemble. For more detail, see Roulston *et al.* (2003).

9.7.2 Ensembles and wave risk

Our second example comes from a research project led by Mark Roulston in cooperation with Jerome Ellepola of Royal Dutch Shell. Shell is interested in waves at offshore platforms all over the world, both fixed rigs (for example, oil well platforms) and floating platforms (Floating Production Storage and Offloading Vessels or FPSOVs). The ECMWF ensemble system includes a wave model (Jansen *et al.* 1997). Our aim is to evaluate the relative ignorance both of dressed ensemble forecasts and the dressed high-resolution forecast. In this case buoy data supplied by Shell provide the target observations and (out-of-sample) verification. The results here differ significantly at Bonga, a floating FPSOV off the west coast of Africa, and at Draugen, a fixed platform in the North Sea. Details of this study can be found in Roulston *et al.* (2005).

At Bonga, we find no statistically significant advantage in using the ensemble forecasts of significant wave height in the frequency bands of interest to Shell, even at day 10. Physically, one might argue that the waves arriving at Bonga now tend to have originated far away in space–time: having a good atmospheric forcing in day 1 and 2 yields a low ignorance wave forecast even at day 10. In that case, the wind that

generates the relevant waves has already 'hit the water' well before the waves reach the FPSOV, suggesting that the current forecasts may contain useful information at Bonga well beyond day 10. Alternatively, one might argue that the current ECMWF ensemble does not target the tropics, reducing the relevance of the ensemble at Bonga, and that doing so would increase the value of the wave ensemble forecast in week one. Either way, after checking for statistically insignificant but economically important extreme events, Shell might argue that there was no reason to buy more than the dressed BFG for Bonga. Of course, it is also possible that increasing the size of the forecast archive might increase the relative skill of the ensemble-based forecast.

At Draugen the situation is quite different, relatively fast-growing nearby weather events sampled in the ensemble result in a significant information advantage for the ensemble wave height forecasts. In the North Sea the probability forecasts based on the ensembles have a clear advantage over those from the high-resolution simulations. These results suggest that significant thought should go into setting the price structure for BFG-based probability forecasts and EPS-based forecasts (and yet other multimodel options which dress the union of BFG and EPS simulations). Such issues are relevant to the economics of forecasting, if not economic forecasts; Charlotte is interested in both.

9.7.3 Electricity demand: probabilistic forecasts or probabilities?

The third example involves forecasting electricity demand in California, a nearly ideal 'how much' question. Modern economies run on electricity; to maintain a reliable electricity grid one must first forecast likely demand and then arrange enough generation to meet that demand. Producing excess generation is expensive, while not having enough generation can be catastrophic (Altalo and Smith, 2004). This asymmetry of impact between positive and negative forecast errors is common in industry; it brings the difficulties of using 'probability' forecasts into sharp focus. A case study for the California electricity grid can be found in Smith *et al.* (2005).

If we do not expect our probability forecasts to be accountable, then we should not be surprised when traditional methods for using these forecasts, such as maximising the expected utility, fail miserably in practice. It is clear that we have extracted useful probabilistic information from our multi-initial condition, multimodel ensembles; it is not at all clear that from this information we can extract a probability forecast which is useful as such. A method to do so would be of great interest scientifically, of great value economically, and of great use socially. In the meantime, however, alternative ad hoc methods for using these predictive distributions are sometimes found to yield more statistically robust results in practice.

9.8 So what is 'uncertainty in initial condition', really?

This may seem a trivial question. The standard answer to this trivial question would be the equation

$$e = x - \mathbf{X}$$

where e is the uncertainty in the initial condition, x is the model state we used at $t = 0$, and $\mathbf{X}$ is the true state of the real system at time $t = 0$. There is a small difficulty here. Perhaps unexpectedly[15] it is with the symbol '$-$', which is undefined as when contrasting apples and oranges, making our one equation ill-posed. While x sits in our model-state space (without doubt one large integer in this digital age), $\mathbf{X}$ is, at best, the true state of the atmosphere. These vectors, if they both exist, exist in different spaces even if we confuse model-temperature with temperature (see Smith, 1997). There is an easy mathematical way out here, simply replacing $\mathbf{X}$ with $\mathbf{P}(\mathbf{X})$ where $\mathbf{P}$ is some projection operator that takes the true state space into our model-state space (this is touched upon in Smith, 2001, Orrell *et al.*, 2001 and Judd and Smith, 2004). Introducing the projection operator shifts the ambiguity in the minus sign to the projection operator, $\mathbf{P}$. It is not clear that we can nail down $\mathbf{P}$ beyond simply saying that it is many-to-many, which may prove more troubling than it seems at first.

The main difficulty in interpreting this equation, however, may be of a different sort: if there is no model state that shadows the time series of each consecutive analysis to within the observational uncertainty, then the shortcomings of the forecast simply cannot be attributed to uncertainty in the initial condition. Why? Because in this case, there was no model initial condition to be uncertain of. It was not that we did not know which value of x to use, but that there was no value of x which would have given an accurate (useful) point forecast. There is no clear definition of uncertainty in the initial condition outside the perfect model scenario.

Both the projection operator and tests of empirical adequacy are bound up with the definition of observational 'noise', while the identification of shadowing trajectories requires projecting observational noise into the model state space. Although most scientists believe that they can recognise it when they see it, much remains to be said about the concept of noise. Turning to a more practical matter: how might we proceed in practice other than by developing the best model structure our technology can support, inserting physically motivated parametrisations with empirically estimated parameter values, insisting that water freeze at exactly zero degrees, and trusting that in time our model will slowly approach Truth?

There is an alternative. Its implications are not yet clear, but if I am lucky enough to be invited to Predictability 2009 then I hope to learn more at that time. The alternative is to embrace model inadequacy while relinquishing the twenty-first century Laplacian dream of accountable probability forecasts. To adopt instead a goal which is less attractive but conceivably attainable: using ensembles of initial conditions

evolved under a collection of imperfect models, aiming to say as much as is justified by our models. We should expect our forecasts to be blatantly wrong as rarely as possible, but not less rarely than possible.

9.9 Conclusions and prospects

Predictability present: There is no doubt that the current operational ensemble systems have value beyond that recognised in industry; this is an opportunity. The question should be seen as one of how to exploit this information content, not as to whether or not it 'exists'. Each of our three users can base better economic decisions and reduce their exposure to weather risk by using probabilistic forecasts based on existing ensemble prediction systems. This is not to say that current probability forecasts are accountable, but rather that current ensemble forecasts are valuable. Charlie, who deals for the most part with binary decisions, can determine the ideal probability thresholds at which he should act and interpret forecasts more profitably. Charles can translate the forecast probabilities both better to gauge the likely behaviour of weather derivatives in the near term, and to extract the likely impact of weather on the futures markets. And Charlotte, our most numerate user, can incorporate probabilistic weather forecasts in a variety of applications from hours to weeks, with the aim of including probabilistic weather information to allow seamless forward planning for weather impacts. Current ensemble prediction systems have demonstrable economic value.

Two obvious questions arise. First, what can be done to raise the level of exploitation of these forecasts? And second, how can we best move forward to increase their value? The answer to the first question involves education, technology transfer, and both the production and advertisement of case studies illustrating the value of current forecast products in realistic economic examples. Answering the question of how best to move forward would, no doubt, benefit from a better understanding of what constitutes the notion of 'forward', but the improvement of forecast models and their associated data assimilation, the improved generation and retention of members in initial condition ensembles, the wider use of multimodel ensembles and improved methods for translating ensembles of simulations into weather forecasts will each play a role.

Almost a century ago, L. F. Richardson began the first numerical weather forecast by hand, while envisioning the use of parallel computing in numerical weather forecasting. Today, the electronic computer plays two rather distinct roles in physical science. First it allows us to calculate approximate solutions to a wide variety of equations at speeds Richardson could only dream of. Second, and perhaps even more importantly, it allows us to record and access data that, in turn, reveal just how imperfect our models are. Having accepted that probabilistic forecasts are here to stay, it will be interesting to see when it proves profitable to shift from trying to make

our model perfect towards trying to make our forecasts better. Accepting that our best dynamical models are not and never need be empirically adequate will open new avenues toward understanding the physics of the Earth System, and may allow us to achieve predictability past the limitations we face at present.

Acknowledgements

It is difficult to write the acknowledgements of a paper that covers a decade. The more important of these insights were obtained through joint work with Mark Roulston, Kevin Judd, Jim Hansen, David Orrell and Liam Clarke while Jost van Hardenburg supplied critical calculations and insights. Myles Allen, Judy Curry, Milena Cuellar, Tim Palmer, Dave Stainforth, Alan Thorpe and Antje Weisheimer made useful comments on earlier drafts, and I am grateful to numerous others for discussions during and after the ECMWF Seminar. Contributions by Isla Gilmore, Pat McSharry and Christine Ziehmann helped lay the foundations on which this iteration was built, clarified by recent discussions with Jim Berger, Jochen Broecher, Devin Kilminster, Ele Montuschi and Naomi Oreskes.

I happily acknowledge personal and philosophical debts to Nancy Cartwright, Donald Giles and Robert Bishop, each of whom I hope to repay. Also, I again thank Tim Palmer for the invitation to the 1995 Predictability Seminar, which had a significant if unpredictable impact on my research; I hope he does not regret the results. LSE CATS's Faraday Partners, NG Transco, EDF, London Electricity and Risk Management Solutions have contributed to my understanding of economics of weather forecasting; I am particularly grateful to Mary Altalo, Melvin Brown, Neil Gordon, Steve Jewson, Shanti Majithia and Dave Parker. Roulston, Judd and myself have each benefited from the Predictability DRI under grant N00014-99-1-0056. Kilminster and myself have also been supported by the USWRP and NOAA. I gratefully acknowledge the continuing support of Pembroke College, Oxford.

Notes

1. The use of CCGT generators comes from the fact that their efficiency in converting fuel to electricity varies with temperature, pressure and humidity; any generation method with weather dependent efficiency would suffice here. For details on forecasting for CCGT generators, see Gordon and Smith (2005).
2. The construction of state-dependent conditional probability distributions from ensembles requires having enough members to estimate a *distribution* of the $n + 1$st variable, given particular values of the first n variables. This is just another guise of the curse of dimensionality, made worse both by the need for a distribution of the target variable and by Charlotte's particular interest in the tails of (each of the many distinct conditional) distributions.
3. After Laplace, see Nagel (1961). Note that the distinction between uncertainty in initial conditions and model inadequacy was clear to Laplace (although perhaps not to Bayes); Laplace distinguished uncertainty of the current state of the universe from 'ignorance of true causes'.
4. It is not clear how well the Bayesian paradigm (or any other) can cope with model inadequacy, specifically whether or not it can yield empirically relevant probability

forecasts given an imperfect model class. I am grateful to Jim Berger for introducing me to a Bayesian research programme that strives for a systematic approach to extracting approximately accountable probability forecasts in real time.

5. This section has generated such varied and voluminous feedback that I am loath to alter it at all. A few things might be clarified in this footnote. First I realise that ground frost may occur when the 2-metre temperature is well above zero, that the freezing point of fresh water might be less relevant to the Earth System than the freezing point of sea water, and (I've learned) that the freezing point of deliquescent haze is quite variable. Yet each of these variables is amenable to physical measurement; my point is simply that their 40 km resolution model-variable namesakes are not, except inasmuch as they affect the model trajectories. Granted, in the lab 'zero degrees C' may be argued exact by definition; any reader distracted by this fact should read the phrase as '32 degrees F' throughout. I am also well aware that the value given to one parameter will affect others, but I would avoid the suggestion that 'two wrongs make a right': outside PMS there is no 'right'. Whenever different combinations of parameter values shadow, then we should keep (sample) all such combinations until such time as we can distinguish between them given the observations. This is one reason why I wish to avoid, as far as possible, the notion of shadowing anything other than the observations themselves, since comparisons with averages and the like might yield the 'two wrongs' without the 'better'. Lastly I realise that it is *always* possible that out-of-sample, the parameters may fail to yield a reasonable forecast however they have been tuned: all forecasts must be interpreted within the rosy scenario, as discussed in Smith (2002).
6. A kernel is a distribution function used to smooth the single value from each simulation, either by substituting the kernel itself or by sampling from it. Note that different kernels can be applied to different ensemble members, as long as they can be distinguished without reference to any future verification (for example, one would expect the ECMWF EPS control member to have a different kernel and be more heavily weighted than the perturbation members at short lead-times; but even the rank order of members may prove useful at longer lead times).
7. As far as I know, ignorance was first introduced in this context by Good (1952) who went so far as to suggest that the wages of British meteorologists be based on the score of their forecasts.
8. In each round, the stake is divided across all options, the fraction on each option proportional to the predicted probability of that option. Kelly (1956) was in fact interested in interpreting his result in terms of information theory; see Epstein (1977).
9. A typical definition of 'fair odds' would be those odds on which one would happily accept either side of the bet; it is not clear that this makes sense outside PMS. Operational fair odds should allow the house offering those odds some opportunity of remaining solvent even if it does not have access to some non-existent perfect probability forecasts. One might define fair odds outside PMS as those set by a not-for-profit house aiming only to maintain its endowment while providing a risk management resource (L. A. Smith *et al.*, unpublished data).

10. A necessary condition for 'objective' as used here is that the forecasts are accountable. This can be tested empirically; we have found no dynamic physical system for which accountable forecasts are available. This is a stronger constraint than other useful meanings of objective probabilities; for example that all 'rational men' would converge to the same probability forecast given the same observations and background information. Subjective probability forecasts can, of course, be obtained quite easily from anyone.
11. The phrase *out-of-sample* reflects the situation where the data used in evaluation were not known before the test was made, in particular that the data were not used in constructing the model or determining parameters. If the same data were used in building the model and testing it, then the test is *in-sample*. Just as passing an in-sample test is less significant than passing an out-of-sample test, failing an in-sample test is more damning. Note that data are only out-of-sample once.
12. Our theories often involve variables that cannot be fully observed; philosophers would call variables like a temperature field *theory-laden*. Arguably, this temperature field is something rather different from its *in silico* realisation in a particular computational model. Thus the model variables composing, say, the T42 temperature 'field' might be called *model-laden*.
13. I would have said no chance, but Myles Allen argues effectively that climate models *might* provide information on 'climate variables' without accurately resolving detailed aspects of the Earth System.
14. 'Post-what?' one might ask, as a computer simulation is not a forecast until expressed in terms of an observable weather variable.Verification against the analysis may be a necessary evil, but its limitations make it inferior to verification against observations.
15. As with an embarrassing number of insights I initially felt were unexpected, Ed Lorenz provided a clear discussion of this issue several decades ago; in this case, for example, he explicitly discussed models that were similar enough to be 'subtractable' (Lorenz, 1985).

References

Allen, M. R. (1999). Do-it-yourself climate modelling. *Nature*, **401**, 642.

Altalo M. G. and L. A. Smith (2004). Using ensemble weather forecasts to manage utilities risk. *Environ. Finance*, **20**, 8–9.

Angstrom, A. K. (1919). Probability and practical weather forecasting. *Centraltryckeriet Tecknologforeningens Forlag*.

Beven, K. J. (2002). Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system. *Hydrol. Process.*, **16(2)**, 189–206.

Bishop, R. (2003). On separating prediction from determinism. *Erkenntnis*, **58**, 169–88.

Borel, E. (1950). *Probability and Certainty.* New York: Walker.

Buizza, R., Miller, M. J. and T. N. Palmer (1999). Stochastic simulation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy Meteor. Soc.*, **125**, 2887–908.

Earman, J. (1986). *A Primer on Determinism.* D. Reidel.

Epstein, R. A. (1977). *The Theory of Gambling and Statistical Logic*. Academic Press.

Fraassen, B. C. van (2002). *The Empirical Stance*. Yale University Press.

Gillies, D. (2000). *Philosophical Theories of Probability*. Routledge.

Gilmour, I., L. A. Smith and R. Buizza (2001). Linear regime duration: is 24 hours a long time in synoptic weather forecasting? *J. Atmos. Sci.*, **22**, 3525–39.

Good, I. (1952). Rational decisions. *J. Roy. Stat. Soc. B*, **14**.

Gordon N. and L. A. Smith (2005). Weather forecasting for combined Cycle Gas Turbine (CCGT). *Proceedings of the First THORPEX Science Symposium* (in press). American Meteorological Society.

Hansen, J. A. and L. A. Smith (2000). The role of operational constraints in selecting supplementary observations. *J. Atmos. Sci.* **57**(17), 2859–71.

Hansen, J. A. and L. A. Smith (2001). Probabilistic noise reduction. *Tellus*, **5**, 585–98.

Janssen, P. A. E. M., B. Hansen and J. R. Bidlot (1997). Verification of the ECMWF wave forecasting system buoy and altimeter data. *Weather Forecast.*, **12**, 763–84.

Judd, K. (2003). Nonlinear state estimation, indistinguishable states and the extended Kalman filter, *Physica D*, **183**, 273–81.

Judd, K. and L. A. Smith (2001). Indistinguishable states. I: The perfect model scenario. *Physica D*, 125–41.

(2004). Indistinguishable states. II: Imperfect model scenario. Physica *D*, **196**, 224–42.

Kelly, J. L. Jr. (1956). A new interpretation of information rate. *Bell System Technical J.*, **35**(4), 917–26.

Kennedy, M. and A. OHagan (2001). Bayesian calibration of computer codes. *J. Roy. Stat. Soc. B, 63*, 425–64.

Kleeman, R. (2002). Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, **50**, 2057–72.

Krishnamurti, T. N., *et al.* (1999). Improved skills for weather and seasonal climate forecasts from multimodel superensemble. *Science*, Sept 3.

Lorenz, E. N. (1985). The growth of errors in prediction. In *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, ed. M. Ghil, pp. 243–65. North Holland.

Lorenz, E. N. and K. Emanuel (1998). Optimal sites for supplementary weather observations. *J. Atmos. Sci.*, **55**, 399–414.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Weather Rev.*, **105**, 803–16.

Nagel, E. (1961). *The Structure of Science*. Harcourt, Brace and World.

Orrell, D., L. A. Smith, T. Palmer and J. Barkmeijer (2001). Model error and operational weather forecasts. *Nonlinear Proc. Geoph.*, **8**, 357–71.

Palmer, T. N. (2001). A nonlinear perspective on model error. *Quart. J. Roy. Meteor. Soc.*, **127**, 279–304.

(2002). The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–74.

Popper, K. (1956). *The Open Universe*. Routledge. Reprinted 1982.

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–68.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.*, **130**(6), 1653–60.

(2003). Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30.

(2004). The boy who cried wolf revisited: the impact of false alarm intolerance on cost-loss scenarios. *Weather Forecast.*, **19**(2), 391–7.

Roulston, M. S., C. Ziehmann and L. A. Smith (2001). *A Forecast Reliability Index from Ensembles: A Comparison of Methods*. Report. Deutscher Wetterdienst.

Roulston, M. S., Kaplan, D. T., Hardenberg, J. and Smith, L. A. (2003). Using medium range weather forecasts to improve the value of wind energy production. *Renewable Energy*, **28**(4), 585–602.

Roulston, M. S., J. Ellepola, J. von Hardenberg and L. A. Smith (2005). Forecasting wave height probabilities with numerical weather prediction models. *Ocean Eng.*, **32**, 1841–63.

Smith, L. A. (1996). Accountability and error in forecasts. In *Proceedings of the 1995 Predictability Seminar*, ECMWF.

(1997). The Maintenance of Uncertainty. In *Proceedings of the International School of Physics (Enrico Fermi)*, Course CXXXIII, ed. G. Cini (Societa Italiana di Fisica, Bologna), pp. 177–368.

(2001). Disentangling uncertainty and error: on the predictability of nonlinear systems. In *Nonlinear Dynamics and Statistics*, ed. A. I. Mees, pp. 31–64, Boston: Birkhauser.

(2002). What might we learn from climate forecasts? *Proc. Nat. Acad. Sci.*, **99**, 2487–92.

Smith, L. A., M. Roulston and J. von Hardenberg (2001). *End to End Ensemble Forecasting: Towards Evaluating the Economic Value of the Ensemble Prediction System.* Technical Memorandum 336. ECMWF.

Smith, L. A., M. Altalo and C. Ziehmann (2005). Predictive distributions from an ensemble of weather forecasts: extracting California electricity demand from imperfect weather models. *Physica D* (in press).

Stainforth, D. A., T. Aina, C. Christensen, *et al.* (2005). Evaluating uncertainty in the climate response to changing levels of greenhouse gases. *Nature* **433**(7024), 403–6.

Weisheimer, A., L. A. Smith and K. Judd (2005). A new view of seasonal forecast skill: bounding boxes from the DEMETER ensemble forecasts. *Tellus*, **57A**, 265–79.