In: Nonlinear Dynamics and Statistics, A. Mees (ed.), Birkhauser, 2000

Chapter 2

Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems

Leonard A. Smith

ABSTRACT
Chaos places no a priori restrictions on predictability: any uncertainty in the initial condition can be evolved and then quantified as a function of forecast time. If a specified accuracy at a given future time is desired, a perfect model can specify the initial accuracy required to obtain it, and accountable ensemble forecasts can be obtained for each unknown initial condition. Statistics which reflect the global properties of infinitesimals, such as Lyapunov exponents which define "chaos", limit predictability only in the simplest mathematical examples. Model error, on the other hand, makes forecasting a dubious endeavor. Forecasting with uncertain initial conditions in the perfect model scenario is contrasted with the case where a perfect model is unavailable, perhaps nonexistent. Applications to both low (2 to 400) dimensional models and high (10^7) dimensional models are discussed. For real physical systems no perfect model exists; the limitations of near-perfect models are considered, as is the relevance of the recurrence time of the system in terms of the likely duration of observations. It is argued that in the absence of a perfect model, a perfect ensemble does not exist and hence no accountable forecast scheme exists: accurate probabilistic forecasts cannot be made even when the statistics of the observational uncertainty are known exactly. Nevertheless, ensemble forecasts are required when initial conditions are uncertain; returning to single best guess forecasts is not an option. Both the relevance of these observations to operational forecasts and alternatives to aiming for exact probabilistic forecasts are discussed.

2.1 Introduction

> All my means are sane, my motive and my object mad.
> Captain Ahab [42]

This paper discusses the limits that uncertainty in the initial condition and error in the model place both on individual forecasts and on predictability in general. The systems of interest will be nonlinear, potentially

chaotic. The methods of analysis and means of computation are sane, and may be assumed exact without altering the limits discussed below. The issue is rather whether or not our questions are well-posed: is the object of our search unobtainable even in the best of circumstances?

It has long been known (see, for example, Brillouin [12]) that even in a well-understood and accurately examined physical system, the combination of *observational uncertainty* and *model error* places severe limits on what we can say about the future of the system. While the remarks below hold for systems as simple as an analog circuit, they will be interpreted in the jargon of weather forecasting, even though the Earth's atmosphere/ocean system is not particularly well observed, nor are current models near-perfect. Nevertheless, numerical weather prediction (NWP) is an appropriate choice since, due to its economic importance, operational forecasts must be made every day and a great deal of thought has gone into attempting to improve the forecasts using any means available. Unlike the armchair forecasts of nonlinear dynamics or theoretical economics, operational weather forecasters must face their failures. Daily. This led Thompson [63] to contrast the relative contributions of uncertainty in the initial condition and model error in the 1950s. In 1965, variations in the reliability of individual forecasts led Lorenz [38] to suggest one (now operational) approach to quantifying the likely impact of uncertainty in initial condition on each particular forecast. Shortly thereafter, Epstein [16] and Leith [32] investigated both computational and analytic limits to maintaining initial uncertainty throughout a forecast. Many issues of current interest to nonlinear dynamicists are old chestnuts of the weather forecasting community.

For many years now, operational centers have made ensemble forecasts: a collection of initial conditions, each consistent with the observational uncertainty, are integrated forward in time. The role of uncertainty is introduced in Section 2.2. In Section 2.3, ensemble forecasting is explored within the perfect model scenario, and some jargon normalization is provided. The ensemble approach to forecasting deterministic systems replaces the single "best guess" initial condition of the traditional approach with a relatively small ensemble of different initial conditions, each member of the ensemble being consistent with the observational uncertainty in the initial state of the system. The idea here is that any initial uncertainty in the initial condition is reflected in the evolution of the ensemble, which in turn reflects the importance of that uncertainty in today's forecast. By observing how quickly the ensemble spreads out (or shrinks), one obtains a local estimate of the stability of forecasts made in this region of the system's state space; global measures like Lyapunov exponents are useless here [59, 57] except in the most simple, uniform systems. Even localized Lyapunov exponents [38, 3, 67] are misleading [70, 60], since they are based on the linearized dynamics over a pre-defined period of time, while the ensemble members may well sample the relevant nonlinearities and indicate when it is that they appear. Indeed, chaos places no *a priori* limits on predictabil-

ity: given a perfect model, ensembles can accurately reflect the likelihood of observing various future conditions (*i.e.,* provide an series of accountable probability forecasts). Such ensembles will slowly evolve towards the invariant measure of the system; but the time scale on which this happens is independent of the measures used to define chaos which are, in turn, based upon the statistics of infinitesimals. Since there is always uncertainty in the initial condition, all nonlinear forecasts should be ensemble forecasts, and the issues discussed below should find application to low dimensional chaotic systems as well as high dimensional weather forecasts.

The stated aim of ensemble forecasts ranges from estimating the ideal forecast probability density function (PDF) to simply obtaining a rough guide to the reliability of today's "best guess" or the control forecast. While the second aim remains in sight, the first cannot be fully realized. A major conclusion of this paper is that just as uncertainty in the initial condition severely limits the utility of a single forecast even in a perfect model, so model error severely limits attempts to obtain "the" forecast PDF. This clarifies the limited applicability of results drawn from within the perfect model scenario. All models are wrong. Yet some are more useful than others. If imperfect models are judged by a standard which they cannot achieve, then the more useful models may be discarded. A similar situation holds when judging between single forecast models by using forecast error: even a perfect model of a chaotic system will have a larger forecast error than a model which predicts the observed mean, at least in the far future. Predicting the mean may be desirable, if one really wants to minimize single forecast error, but this approach is obviously a poor guide to improving the physics of the model.

A basic difficulty in evaluating ensemble forecasts comes from the fact that the ensemble forecast estimates a probability density function in state space, while the verification (the true state of the system at the forecast time) is a point in state space[1]. It is not possible to verify a single probability forecast, and each forecast involves a different initial condition. Further, no two initial conditions will ever be close in a dynamical system where the time required for the system to return to a point near the current state (*i.e.,* the recurrence time) is longer than the likely duration of observations; thus the details of each PDF will differ for each forecast. The evaluation of a series of probability forecasts, given that each forecast PDF is different and that only a single realization of each forecast exists, is discussed in Section 2.4, where the one-dimensional method due to Talagrand is generalized to higher dimensional spaces. But once it is accepted that an accurate forecast PDF cannot be obtained even in near perfect models, then new

[1]Worse still, there are at least three relevant spaces here: forecasts lie in the model-state space, the system lies in the "true" state space, and observations explore yet another.

methods both of inter-model comparison and of multi-model forecasts are called for; this may prove especially important in guiding model development. After a realistic look at the ambiguities introduced by model error in Section 2.5, two alternatives to computing a forecast PDF are introduced: (i) aiming for a bounding box, and (ii) aiming for a ϕ-shadowing orbit. Each of these can be used to determine admissible predictability times. Fully embracing the limitations discussed below suggests a new method for combining (rather than selecting the best of) imperfect models: the cross pollination in time (CPT) ensemble strategies introduced in Section 2.6 can outperform all of the models available in terms of the two aims stated above. Standard multiple model inter-comparisons search for the best model in the same way that standard data assimilation routines search for the true state of the system; if no unique state can be identified empirically even under ideal conditions, then there is no "true" state, and each of these standard approaches may hinder the resulting forecast. This holds regardless of how sane and sophisticated the techniques employed in the endeavor may be.

2.2 Uncertainty

Consider an intelligence which knew all the laws of nature precisely, and had accurately (but not exactly) observed an isolated chaotic system for an arbitrarily long time. Such an agent - even if sufficiently vast to subject all this data to computationally exact 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. While our agent could not predict the future precisely, the future would hold no surprises for her: the predictability of the current "state" she could see [28, 56]. By forming an ensemble forecast from the plausible initial conditions consistent with both the system and the observations, she could estimate the probability density function (PDF) of future states to any desired accuracy. And these ensemble forecasts would be accountable: as the number of members in the ensemble grew, the accuracy of the PDF would improve proportionately. Further, for each particular initial state, she could specify the accuracy of observation required to allow a desired level of accuracy in the final state [53, 56, 57]. She has not only a perfect model, but also a perfect ensemble: a set of initial conditions both consistent with all observations and "on the attractor." The true trajectory can be viewed as just another member of the distribution which she samples to form the ensemble.

Operational forecasters at major weather centers in both Europe and North America, attempt an impersonation of this intelligence daily when they perform ensemble forecasts (see Palmer *et al.* [50], Toth and Kalnay [65], and references thereof). The predictability of the atmosphere varies from day to day, and so a single "best guess" forecast is incomplete without a

daily estimate of its likely accuracy. Ensemble forecasts aim to foresee variations in predictability by quantifying the time required for a given day's ensemble members to splay out along significantly different trajectories, thereby quantifying the point at which that day's "best guess" forecast is unlikely to be accurate. Ideally, one could also use the ensemble to quantify the probability of various events. But no physical model is perfect, and as we shall see, model error may make accountable probability forecasts unreachable, just as observational uncertainty makes a single forecast of little value. Our agent achieves an accountable forecast by evolving a perfect ensemble under a perfect model; once imperfect models are in use, no perfect ensemble exists. Accepting this fact forces us to change the interpretation and goals of forecasts. In fact, it calls into question what is meant by the state of a physical system.

Traditionally, the current state of a deterministic system is regarded as a point in state space, the exact location of which is obscured by observational uncertainty. This scenario only arises in computer experiments where we determine a trajectory and then pretend to forget where it was after adding some simulated observational noise. Even in that case, given only the noisy observations it would not be possible to identify a true state if we did not already know the answer: there would be a range of initial conditions, parameter values, and even distinct model structures which provided equally valid descriptions of the data. Clearly the traditional notion of "the state" of the system must be empirically suspect if even our idealized agent could not identify this "state" given a perfect model. In reality, of course, all models are wrong. It is our models which have states; there is no need for the hypothesis that physical systems do.

2.3 The Perfect Model Scenario

What is the perfect model scenario? Let the role of the physical system be played by a set of equations proposed by Lorenz [37] as a parody of some atmospheric variable. As shown schematically in Figure 1, the system consists of m slow large-scale variables (the $\tilde{x}_i$) and $m \times n$ fast small-scale variables (the $y_{i,j}$) and thus has a state space dimension of $m(n+1)$. The notation $\tilde{x}$ is used to distinguish variables in the system state space from those in the model-state space, which will be denoted as x. Details can be found in Lorenz[37], Hansen [24], Orrell [48], Hansen and Smith [25] and the references therein. The equations are:

$$\frac{d\tilde{x}_i}{dt} = -\tilde{x}_{i-2}\tilde{x}_{i-1} + \tilde{x}_{i-1}\tilde{x}_{i+1} - \tilde{x}_i + F - \frac{h_{\tilde{x}}c}{b}\sum_{j=1}^{n}\tilde{y}_{j,i} \qquad (2.1)$$

$$\frac{d\tilde{y}_{j,i}}{dt} = cb\tilde{y}_{j+1,i}\,(\tilde{y}_{j-1,i} - \tilde{y}_{j+2,i}) - c\tilde{y}_{j,i} + \frac{h_{\tilde{y}}c}{b}\tilde{x}_i. \qquad (2.2)$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$ and with cyclic boundary conditions on both the $\tilde{x}_i$ and the $\tilde{y}_{j,i}$ (that is $\tilde{x}_{m+1} = \tilde{x}_1$, $\tilde{y}_{(n+1,i)} = \tilde{y}_{(1,i)}$ and so on). In the computations below $F = 10$, $m = 8$ and $n = 4$. The constants b and c are both equal to 10, so the small-scale dynamics are 10 time faster (and a factor of 10 smaller) than the large-scale dynamics, while the coupling coefficients $h_{\tilde{x}}$ and $\tilde{y}$ are both set to unity.

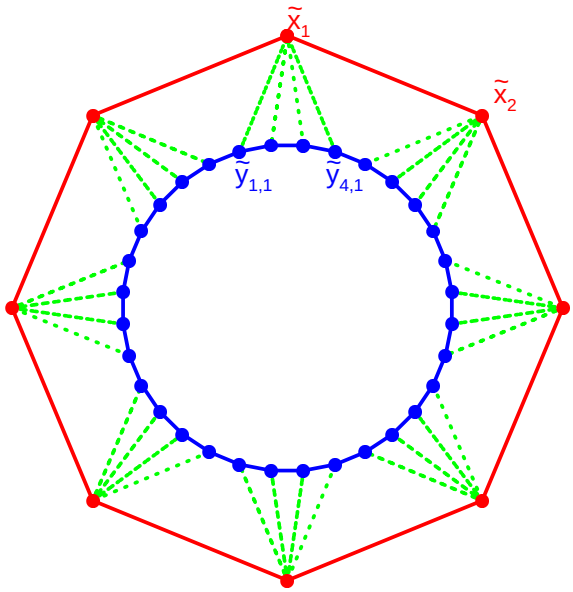

FIGURE 2.1. Schematic of the Lorenz two-scale system.

Now some jargon.

When the forecast model is used to generate the observations which are in turn to be forecast, one is in the *perfect model scenario*. The actual state of the system will be called *truth*, while our best estimate of that state, given only limited, noisy observations, is commonly referred to as the *analysis*. To test our model, the forecast is contrasted with the *verification*, which is in practice a future analysis; in (and only in) a perfect model experiment can the verification really be truth itself. For a single set of simultaneous observations, the uncertainty in the analysis is related to the observational uncertainty. Given a time series of observations, the

analysis corresponds to our best guess at the state, since this uses all the available observations (and a model), the analysis uncertainty in this case may be much lower than the observational uncertainty in the individual measurements. In the perfect model scenario, the analysis uncertainty is less than the observational uncertainty.

Operationally, an analysis may be generated via a 4-dimensional variational assimilation (4DVAR) technique [62, 52]. 4DVAR attempts to locate the free running model trajectory which minimizes the difference between the model trajectory and the observations over a given duration (called the assimilation window), while allowing the observations to be spread out in both space (3-D fields) and time (+1-D). Achieving this in real-time with disparate data sources, each of which has different observational uncertainties and which intermittently vanishes, is nontrivial. The search for a solution is also hampered by local minima in a 10^7 dimensional space, but the key point here is that the resulting analysis can be much more accurate than the measurement uncertainty in a single set of simultaneous observations *as long as* the model is sufficiently accurate. We shall quantify "sufficiently accurate" below, here we note that this approach searches for "the" true state; this is somewhat troubling if we have accepted that there is no unique solution even within the perfect model scenario. An alternative approach to generating a best guess analysis and then creating ensemble members by adding perturbations is to generate an ensemble directly. This approach has been illustrated in simple low dimensional models [28] while an operational method based on multiple analyses has been investigated by Houtekamer *et al.* [26]. Issues surrounding what makes the best analysis or the best operational ensemble are widely debated within the atmospheric community; many other options exist [7, 9, 22, 23, 43].

Traditionally, a weather forecast consisted of a single trajectory, started at the analysis and run at the highest available resolution. Such a traditional "best guess" is often run alongside an ensemble forecast, but since it is run at higher resolution, it lies in a different model-state space from that of the ensemble members. The *control* forecast is the ensemble member starting at the current analysis. Typically, roughly equal computational resources are invested in constructing the analysis and running the ensemble, with the high resolution run taking up most of the remainder ($\approx 10\%$). Open questions include the issue of whether additional computational resource should go towards increasing the model resolution at which the ensemble members are run (*i.e.,* obtaining a better PDF), or running more ensemble members at the current resolution (*i.e.,* a better approximation of an inferior PDF), or running the current system further into the future. Alternatively, resources could be directed towards obtaining a better analysis. This could be approached either through a more computationally intensive assimilations technique, or through obtaining additional observations, the

locations of which may change[2] daily [24, 25, 34].

2.3.1 Forecasting with a Perfect Model

Figure 2.2 shows three ensemble forecasts in the perfect model scenario: A new ensemble is initiated every four time units (as denoted by the circle superimposed upon truth). Although the initial condition is not known exactly, we will assume a perfect model in this section: that is, the equations and parameter values are known exactly and the same integration scheme is used by the model and the system. Further we assume that the system is chaotic, although this assumption is not necessary if we have only a finite duration of observations.

Brillouin [12] clearly shows how observational uncertainty limits our knowledge both of the current state and of the future; general arguments [28] establish that the current state is often not uniquely defined given uncertain observations over any duration. A simple way to see that this is true is to consider a special case of a chaotic dynamical system for which stable and unstable manifolds of the current state exist and where the observational uncertainty is due only to quantization (*i.e.,* truncation error). Clearly, there are portions of the unstable manifold within the current quantization box, which are also in the same series of previously observed boxes; that is, a set of trajectories which agree with all previous observations exactly, say, equal in the first three digits. This implies an infinity of states consistent with the observations. Thus no unique current state is defined by the observations, and therefore there can be no unique future state. Accountable forecasts must consider this infinity of states and attempt to maintain the initial uncertainty, quantifying its evolution during the forecast.

The forecast approach shown in Figure 2.2 will fail in this aim. The perfect model is used, and the initial conditions used are consistent with the uncertainty in the current observation. Since the model is perfect, the ensemble may contain trajectories which remain indistinguishable from the observations arbitrarily far into the future; such a model is said to ι-shadow the system [20, 56, 58]. Further the forecast PDF is a valid Monte Carlo approximation of the Fokker-Plank equations, given the observational uncertainty. In what way then is the forecast PDF incorrect?

When making ensemble forecasts we can estimate the probability of future events simply by counting the number of ensemble members in which the event occurs; counting, for example, the number of ensemble members in which there are clear skies over Oxford for a 24 hour period of interest. By grouping together various forecasts (made on different days) which happen to have the same predicted probability, we can determine the relative

[2]The idea being to take data at locations where the current level of uncertainty most hinders the forecast at some future time.

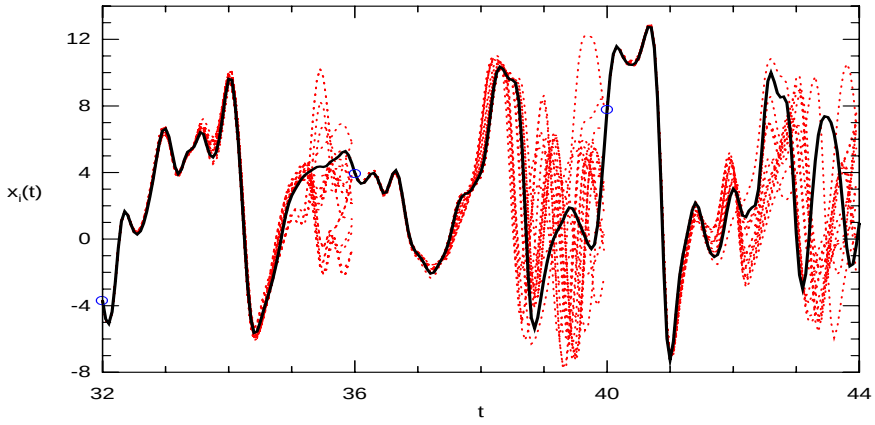

FIGURE 2.2. Perfect model ensemble forecasts for the Lorenz system of Equations 2.1 and 2.2 showing an $\tilde{x}_i$ component of the true trajectory (solid) and of the forecast trajectories (dashed) from three perfect model ensembles. The members of each ensemble are consistent with the initial observational uncertainty. In this case the model and the system are identical but the values of the $\tilde{x}_i$ are imperfectly known; for convenience, the true $\tilde{y}$ values are used in each case. Every four steps an ensemble of initial conditions is forecast (each initiation is denoted by a circle). Visually, one can identify the time at which any one best guess forecast is likely to become unreliable. Yet one cannot obtain an accurate probability forecast from these ensembles, since the probability that an initial state is mistaken of the true state differs from the probability that that state is the true state.

frequency with which the event occurred on the days where the predicted probability was, say, about 10%. Ideally, this relative frequency should be near 0.10. To achieve this ideal requires a model capable of producing a realistic trajectory and an initial ensemble which gives the correct relative weight to physically relevant points consistent with the observational uncertainty. Of course, evaluating the accuracy of extremely low probability events, like the example above, may require extremely long data sets in order to collect enough statistics to obtain a reliable estimate of the relative frequencies (see Murphy [44, 45] and references therein; Smith [57] provides a low dimensional dynamicist's point of view and examples.).

Obviously a perfect model contains initial conditions consistent with the observational uncertainty which ι-shadow for an arbitrarily long time. This is not the question, however. The difficulty lies in determining the subset of initial conditions which are physically relevant. Suppose, that the system evolves on a manifold with dimension less than that of the system state space. Physically relevant points are restricted to the manifold, while the observational uncertainty will, in general, extend into the full state space: we are required to select only initial conditions from a set of zero measure

on a manifold we do not know *a priori* . In this case the probability that a state $\mathbf{y}$ cannot be distinguished from the true state is not equal to the probability that it is the true state; the true state will lie on the manifold, $\mathbf{y}$ need not, as illustrated in Figure 2.3. If we do not restrict our ensemble members to this manifold, then the predicted probabilities will not match the relative frequencies; this is nicely demonstrated by an example due to Gilmour [20], shown in Figure 2.4. The evolving probability distributions in the left column (Figure 2.4 a) reflect ensembles consistent with the analysis uncertainty but *not* constrained to lie on the attractor, as in Figure 2.2. Contrast the unconstrained ensembles in the left column for $1 < t < 2$ and $2 < t < 3$, with the corresponding perfect ensembles in the right column: in each case the unconstrained forecast grows much too wide, due to including initial conditions which cannot be distinguished from the true state given the observations, but also cannot be the true state since they are inconsistent with the (unknown) longterm dynamics (*i.e.,* they are not on the attractor). Interpreted in terms of the schematic in the right panel of Figure 2.3, the unconstrained ensembles choose members from the 2 dimensional plane weighted by the isopleths, while the perfect ensemble only admits points on the attractor (the dots), again weighted by their relative likelihood given the isopleths of uncertainty. The unconstrained ensembles succeed in giving a general estimate of when the (unconstrained) analysis will become unreliable, but unlike the perfect ensemble these unconstrained ensembles cannot yield accountable probability forecasts.

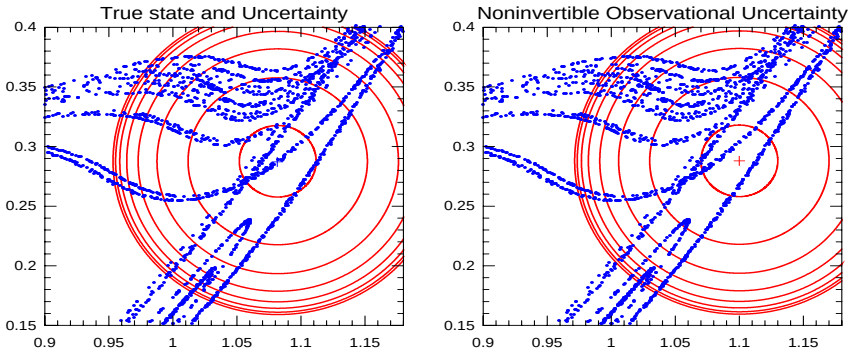

FIGURE 2.3. Left: isopleths of the probability of an observation, given that the true state (+) is (1.0818408, 0.28764392), note that (+) lies on the attractor. Right: isopleths of the probability that a state would give rise to the observation (+) given the observational uncertainty; but without knowing whether a point in state space is on the attractor, one cannot compute the probability of its being the true state. In this panel, the observation (+) is *not* on the attractor.

As indicated on the left panel of Figure 2.3, we can compute the prob-

ability of an observation $\mathbf{x_{obs}}$ given the true state and the statistics of observational uncertainty (or the analysis uncertainty if any noise reduction is attempted). And as indicated on the right panel, we can compute the probability that a given point $\mathbf{x}$ could give rise to the observation $\mathbf{x_{obs}}$, but we **cannot** compute the probability that a given $\mathbf{x}$ is the true state given *only* the observation and the noise process if the system evolves on a lower-dimensional manifold. To obtain the probability that $\mathbf{x}$ is the true state requires[3] additional information (the manifold or the invariant measure in the case of a strange attractor). Without this additional information, the initial PDF will assign positive probability to points in state space which cannot correspond to the true state, and thus the initial PDF will be incorrect. And if the initial PDF is wrong, then the final PDF is wrong, almost certainly. We may be able to state approximately the probability of falling outside a region of state space, but we cannot obtain an accurate probability forecast. This again emphasizes that it is misleading to think of "uncertainty in the initial condition" in terms of a single well-defined state to which a random variable is added to yield the analysis. It is often better to think of "truth" as a random choice from the physical states consistent with the observations. To obtain a perfect ensemble (one with accurate predicted probabilities) one must choose ensemble members from the same distribution with the same relative weighting. In general, if the system is evolving on a lower dimensional manifold (or attractor) this *cannot* be done; at least not without a perfect model and a huge computational effort.

We return to that point in a moment; but first stress that there is nothing "low dimensional" about this manifold: in the 10^7 dimensional systems common in NWP, a $10^7 - 1$ dimensional manifold counts as lower dimensional. Further, many practical forecasting systems (including NWP) are likely to fall into a Catch 22: if the system evolves on a lower dimensional manifold, then obtaining perfect ensembles may prove intractable; but if the number of active degrees of freedom is equal to the dimension of the state space, then there is an insufficient number of observations to initialize the model in the first place. In practice, models can be initialized given the observations, so physical constraints implicit in the equations of the model must lower the effective number of degrees of freedom; but if the system evolves on a lower dimensional manifold then ....

Of course, high dimensional modeling (*e.g.,* those with high spatial resolution) *assumes* that "the physics" restricts the effective number of active degrees of freedom. In practice, weather models tend to evolve the equations of motion of a fluid in a three-dimensional space (either in a grid point form, a spectral form, or both); given this restriction on model structure a

[3]The extent to which this is relevant to NWP is discussed in Section 2.4.2, it is clearly relevant to forecasting systems whenever the the projection of the attractor (or manifold) into the model-state space is lacunar on the length-scales defined by the observational uncertainty. Stephenson [61] notes implications this holds for quantifying analysis error.

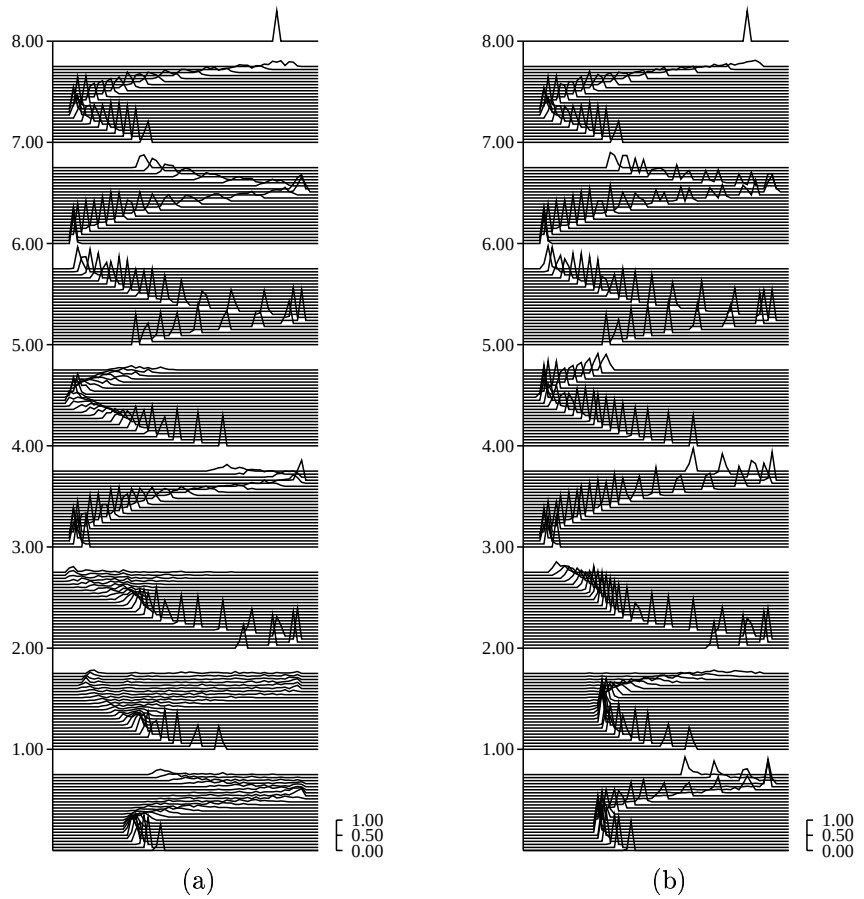

FIGURE 2.4. Comparison of (a) unconstrained ensembles and (b) perfect ensembles based upon the same observations. Each 64 member ensemble is evolved under a perfect model of the Marzec Spiegel system [40] and projected onto $x \in [-1, 1.5]$. Time increases upwards. The gaps in the vertical indicate when new ensembles have been formed about the corresponding observation. Note that the distributions of the perfect ensembles just prior to the gaps tend to be tighter and more closely aligned with the distribution just after the gap (*i.e.*, in closer agreement with the verifying observations). The distribution of the observational uncertainty was $\mathcal{U}_3(0.01)$. Figure from Gilmour [20].

high-resolution model may be required to obtain a good representation of a low-order manifold. Ideally we might use a (lower order) model structure whose model-state space consisted only of the manifold, but at present our understanding of the physics is based on a spatial representation of atmospheric fields, and there is not yet sufficient data to construct the desired manifold empirically. We cannot formulate (much less integrate) the physical equations restricted to this manifold. Good low-order behavior may require high-resolution models[4].

There is something of a symmetry here between NWP forecasts and forecasting on strange attractors via delay embedding, as is common in nonlinear dynamics [54]. A major aim of the dynamic reconstruction in delay space is to model only the manifold, or only the lowest dimensional space within which the manifold can be embedded. But once in this low dimensional space, there is no simple way to return to the physical state space of the system, that is, there is no method for interpreting model states in terms of physical variables other than those observed. Even though the dynamics of the reconstruction are diffeomorphic to dynamics in the full state space (on the attractor), an interpretation in terms of physical variables is much simpler in the full state space. In NWP the difficulty is in restricting the "physical variable" model to the manifold, while in delay reconstructions it lies in interpreting points on the manifold in terms of physical variables.

The existence of the right hand column of Figure 2.4 indicates that perfect ensembles are not always unobtainable. Given a perfect model, the issue is one of computational expense which is, in turn, determined by the resolution of the observations and the recurrence time of the system. To build a perfect ensemble, we simply wait for an analog. The relevant question is: how long must we evolve the model before we obtain two states which are indistinguishable given our observational uncertainties? The 64 member perfect ensembles of Figure 2.4 were obtained by collecting analogs in this way [56]. For third-order chaotic systems, this is often computationally feasible; for the Earth's atmosphere, however, a single return to within the current observational accuracy over a large area like the northern hemisphere has been estimated to require 10^{30} years [66]; this is significantly longer than the lifetime of the atmosphere (and likely to exceed that of the Universe, for that matter). And the model *must* be perfect: an arbitrarily good weather model can have a horrid climate, and it is the climate (the attractor) we must sample to obtain good probability weather forecasts. Our agent can do this because she has a perfect model and unlimited computational power. In Section 2.5 we note that if the model is imperfect, no perfect ensembles exist (almost certainly).

---

[4]I am grateful to P. Young and A. Lorenc for persuasively arguing the merits of the low-order approach and of the high-resolution approach, respectively.

We close this section with an epistemological question. In a recurrent system, perfect ensembles can be constructed with an analog approach (assuming that the successive returns are completely decorrelated!); in a non-recurrent system, or a system whose recurrence time is long compared to its likely lifetime, what meaning can be given to an ensemble forecast? Taking uncertainty in the initial condition seriously also raises a practical question: if we hold "truth" to be a point in state space, then we are forecasting a probability distribution in state space which we must verify with a single point. How might we do this?

2.4 Ensemble Verification

For each initial condition, an ensemble of initial states is forecast but only a single state exists with which to verify the forecast[5]. How might we evaluate that ensemble forecast? An individual ensemble forecast cannot be verified, but the consistency of a series of ensemble forecasts can be verified. For forecasts of scalar quantities the standard approach is to use rank histograms [5, 6, 21] commonly refered to as Talagrand diagrams. Assume for the moment that we have a perfect ensemble: our ensemble was chosen from the same distribution as "truth"; in this case nothing can distinguish "truth", it is just another ensemble member. This fact may be exploited, for example, by counting the number of forecasts which are greater than "truth". This is illustrated in Figure 2.5 which shows the evolution of some scalar quantity; time runs from left to right and we have adopted the meteorological technique of denoting the "true" trajectory as a straight horizontal line. Eight member ensembles of model trajectories appear at regular intervals and diverge from "truth" at a rate that depends on the local nonlinear structure of the model. Given a perfect ensemble, the number of ensemble members above "truth", N_{over}, should be uniformly distributed between 0 and N; better still the variance of any one bin in such a histogram is easily estimated. In operational NWP, the first bin and last bin tend to be overpopulated: truth falls outside of the ensemble much too often.

For the imperfect ensembles in the left hand column of Figure 2.4, which are consistent with the observational uncertainty but not constrained to lie on the attractor, the Talagrand diagrams are under-populated at the extremes; this is to be expected when the ensemble regularly contains initial conditions not on the attractor and which diverge rapidly. For the perfect ensembles in the right column of Figure 2.4, the Talagrand diagrams are

[5]One might treat the verification as a PDF consistent with the observational uncertainty and centered upon the analysis, but the results below are easily generalized to that case.

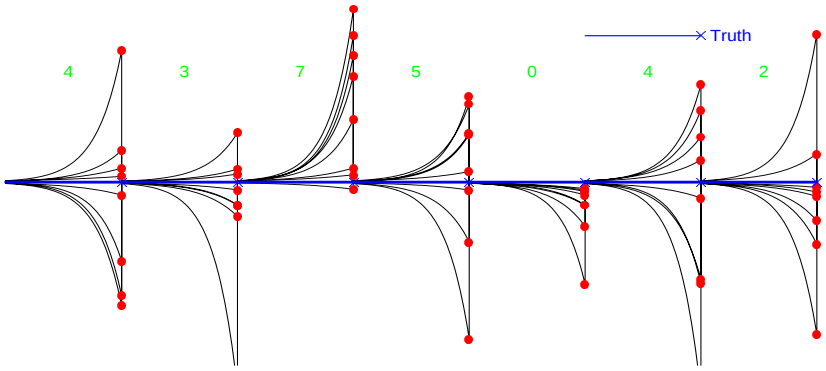

FIGURE 2.5. A schematic of ensemble evaluation one dimension: count N_{over}, the number of forecasts greater than truth for each lead time. If perfect ensembles are used, then N_{over} should be uniformly distributed; in N_{exp} experiments, we expect the relative frequency of a particular value of N_{over} to have mean N_{exp}/N_{bins} and variance $N_{exp}(N_{bins}-1)/N_{bins}^2$, where N_{bins} is just the number of members in the ensemble plus one.

consistent.

Note that the Talagrand diagram can only be used for scalar forecasts since it relies on the rank ordering of the forecast values. Attempting to combine diagrams of different forecast values (say the temperature in London, that in Berlin, and that in Paris; or the geopotential height at each grid points in some region of interest), is ill-advised unless the predictands are truly independent; an unlikely case. Given a perfect ensemble, these combined diagrams would still be flat asymptotically, but we could no longer compute the expected rate of convergence (*i.e.,* the variance), and hence could not determine whether diagrams based on a finite amount of data were consistent with those expected from perfect ensembles or not.

2.4.1 Minimum Spanning Trees

The essence of one-dimensional approach can be generalized to high dimensional spaces by employing minimum spanning trees (MST) [4] to detect whether the ensemble members are simply additional draws from the distribution that generated "truth". The idea is shown in Figure 2.6. Consider a finite set of points in any metric space. A spanning tree is a collection of line segments which connects all the points in a set with no closed loops. The minimal spanning tree is that spanning tree in which the sum of the lengths of the segments is smallest. The MST test then, is to take all N member subsets of the $N+1$ points (the N ensemble members and the

control). If "truth" and the ensemble members are drawn from the same distribution, then no computation can distinguish the spanning tree from which "truth" was omitted [68]: we simply count N_{over}, now the number of the N spanning trees where an ensemble member was omitted whose length is longer than that of the MST where "truth" was omitted. It is not possible to evaluate a single ensemble in this way, but given a collection of n ensemble forecasts, a wide variety of systematic errors in ensemble formation could be identified. Histograms of N_{over} should follow the same statistics as the histograms of the Talagrand diagram, with a relative frequency approaching $\frac{1}{N+1}$ for each of the $(N+1)$ possible results $(0, 1, 2, \ldots, N)$ and variance $\sigma^2 = \frac{1}{n}\frac{N}{(N+1)^2}$, as above[6].

Four examples are shown in Figure 2.7. The upper left panel shows an acceptably flat distribution when both the verification and ensemble members are chosen from the distribution in Figure 2.6. The upper right distribution reflects that when the verification is randomly distributed within the frame of the figure, it is often too far from its nearest neighbor, leading to a small MST when it is omitted, and thus an increasing histogram as shown. Lower left panel shows the histogram which results when each verification is taken from a line lying near the attractor; this graph is easy to reject but its shape is less easy to interpret: again the verification is too often too far from its nearest neighbor, but on those occasions when an ensemble member is chosen from that part of the attractor near the line on which the verification must lie, then the MST length of the tree omitting the verification tends to fall in the middle range. Finally the lower right panel shows the result when the variables are chosen independently, but each from the corresponding correct distribution: the x-component of the verification is taken from a correct distribution of x values and the y-component of the verification is taken from the correct distribution of y values. In this case, both the Talagrand diagram for x and the Talagrand diagram for y would have been found acceptable, but the MST test rejects since the conditional distribution of x given y is incorrect.

2.4.2 Relevance to Operational Forecasting

Hopefully, the last few sections have made clear the difficulty of obtaining perfect ensembles, even given a perfect model. This is without a doubt a concern when forecasting low dimensional systems described by strange attractors; if the perfect ensemble is lacunar and the operational ensemble is not, then accurate probability forecasts will not be obtained. But is this really an issue in operational weather forecasting? In operational forecasts

[6]Note that this is the variance in a given bin over many realizations, since the relative frequency in each bin is not independent (they must sum to one), the variance of the different bins in a single realization will differ from this, particularly when only a small number of bins are used.

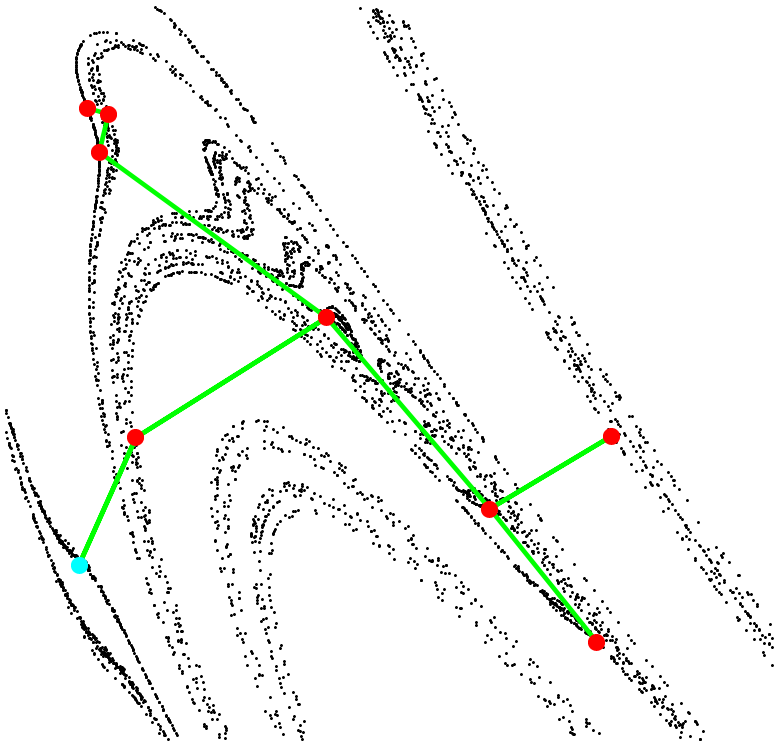

FIGURE 2.6. A minimal spanning tree from the combined set of 8 ensemble members (dark dots) and the verification (light dot) which is also on the attractor (and in this experiment "truth').

where the system is evolving on a lower dimensional manifold or attractor (not low, just lower), and where the structure of the manifold is not isotropic on the length scales resolved by the uncertainty in the analysis, then yes these issues are important. For example, let the true state lie on a line and the analysis uncertainty correspond to a uniform distribution on a disk which the line intersects. In that case sampling the disk to form an ensemble consistent with the analysis uncertainty yields an ensemble very different from the perfect ensemble, which will only contain points from both on the disk *and* on the line. Alternatively, if the manifold consisted of many parallel lines, effectively filling the plane on length-scale defined by the radius of the disk, then the unconstrained ensembles might prove similar to perfect ensembles, as long as they did not contain too many members. In general, the difficulties above may prove less important in systems where the invariant measure is smooth and slowly varying in state space (or its

projection into the model-state space is uniform), or where the manifold is so contorted on the scale of the observational uncertainty that it can be treated as uniform. There may also be cases where the resolution of the model is so coarse as to make the variations unresolvable.

Of course, it is also possible that model error is so large that the forecasts go badly wrong before the effects above come into play. But in the limits of accurate short term prediction models and small uncertainty in the initial condition, the issues above will prove relevant both for low dimensional dynamical systems and for the high dimensional weather models of NWP.

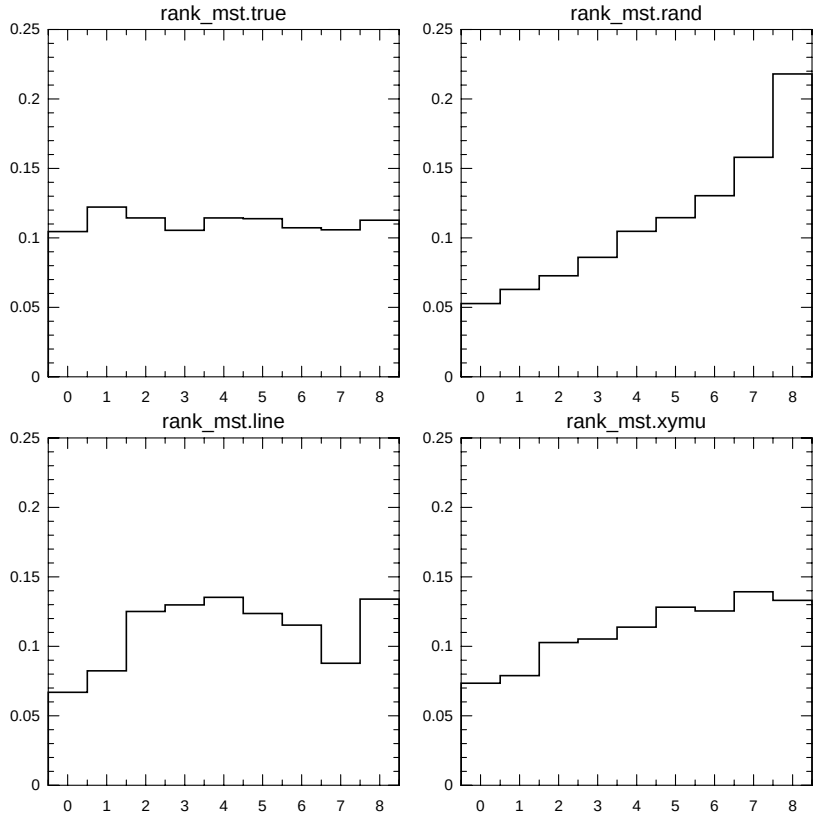

FIGURE 2.7. Each panel shows a histogram of N_{over}, the number of MSTs omitting one ensemble member which were longer than the tree omitting the verification. In every case, the ensemble members were taken from the distribution shown in of Figure 2.6. Histograms reflect when the verification was taken from: the same distribution (top left panel), a uniform distribution in 2-d (top right), a uniform distribution in 1-d (lower left), and with independently chosen x and y components, where the distribution of each component matched that shown in of Figure 2.6 (lower right).

Meteorologists tend to distinguish forecasts made with large models (NWP) from those made using less complicated empirical models and personal insight. While the NWP models get the most press, the simpler methods are sometimes quite good. This is most often true on small spatial scales and short forecast times (hours) at locations for which there are long historical records [71], and on very long time scales (seasonal or greater) where the biases of NWP models may become evident [11, 51]. It would be interesting to contrast the performance of ensembles in these empirical models, with those under NWP for, say, seasonal time scales.

2.5 Imperfect Model Scenarios

Only hypothetical agents are allowed perfect models, we must deal with realistic models. And this fact alters the philosophy of nonlinear forecasting as fundamentally as the acceptance of uncertainty in the initial condition. To see this, we introduce a model for the two level system of the previous section which will play a role analogous to that weather models play in relation to the Earth's atmosphere/ocean system. Keeping Equations 2.1 and 2.2 as the system, we will consider models of the form:

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + P_i(\mathbf{x}, t), \qquad i{=}1, m \tag{2.3}$$

These equations for the model variables $\mathbf{x}$ are structurally similar to Equations 1 which determined the large scale $\tilde{\mathbf{x}}$ dynamics of the system, they differ in that the dynamics of the small scale fast variables, the $\tilde{\mathbf{y}}$, have been parameterized by the function P. A wide range of parameterizations may be entertained; options we have explored for $P_i(\mathbf{x}, t)$ include:

$$P_i(\mathbf{x}, t) \quad = \quad \begin{cases} \alpha_0 & \text{constant} \\ \alpha_0 + \alpha_1 x_i & \text{linear} \\ \alpha_0 + \boldsymbol{\alpha} \cdot \boldsymbol{x} & \text{m-linear} \\ H_1(\mathbf{x}) & \text{nonlocal1} \\ H_2(\mathbf{x}, \frac{\Delta \mathbf{x}}{\Delta t}) & \text{nonlocal2} \\ \text{I.I.D}_{obs} & \text{IID} \\ \gamma_1 P_i(\mathbf{x}, t-1) + N(0, \gamma_0) & \text{AR(1)} \end{cases}$$

These parameterizations range from simple variations on linear models (a constant, a linear parameterization based on only the local variable x_i, a linearization based on all m components[7] of $\mathbf{x}$), through nonlinear

[7] If the x_i's are interpreted as being distributed in physical space, then this last model is nonlocal in physical space since it requires input from other grid points; it is a serious complication given the computational structure of current weather models, but may prove worth the difficulty of implementation as the spatial resolution of those models improves.

variations suggested by prediction studies in low dimensional nonlinear dynamical systems [55, 17, 13, 27, 2, 1] (here H_1 is nonlinear and nonlocal in physical space, while H_2 is also nonlocal in time), and finally to simple stochastic parameterizations (either choosing a value for P_i at random from the observed historical forcing, or fitting an autoregressive model to those observations and using that AR model.).

One property each of these various parameterizations share is that they are wrong: given that $\mathbf{x} \in R^m$ while $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in R^{m(n+1)}$, there is, in general[8], no perfect model with the form of Equations 2.3 and thus no perfect ensemble. Each model will have one distribution from which ensembles may be drawn which will verify at 1-day; and a different distribution yielding ensembles which will verify at 2-days, and so on. And even these distributions will vary from model to model. The forecast quality of each of these models will be discussed elsewhere; the point of introducing them here is to consider the question of what to do with them: should one search for a best model? consider an ensemble over models? or something even more radical? And what is the aim of ensemble forecasting in this context?

We start with an easier question: what is the correct value for α_0 in the constant parameterization? An obvious choice is $\alpha_0 = \tilde{\alpha}$ where

$$\tilde{\alpha} \equiv \left\langle F - \frac{h_{\tilde{x}} c}{b} \sum_{j=1}^{J} \tilde{y}_{j,i} \right\rangle_{\tilde{\mathbf{x}}}, \tag{2.4}$$

that is, the average value of the forcing term where the average is taken over the invariant measure of the true system. But in a nonlinear system, this value has no special claim to optimality; why not take the value which minimizes the one-step forecast error? or minimizes the two-step forecast error? or yields the longest mean ι-shadowing time? or best reproduces the invariant measure of the true attractor projected into the model-state space [41]? For the model-state space differs from the state space, $\mathbf{x} \neq \tilde{\mathbf{x}}$ even if both variables are called x. Thus the correct method for determining the free parameters in the imperfect models above depends on the goal of the forecaster. There need be no unique set of "true' parameter values; standing water need not at freeze at exactly zero degrees C in a good weather model.

In a perfect model, there is a unique perfect ensemble corresponding to all potential states of the system, each weighted by its probability given the observations. In an imperfect model no perfect ensembles exist, and it is doubtful whether a unique optimal ensemble is well-defined for the same reasons that optimal parameters are not. None of these models will

[8]Of course the inclusion of parameterizations H_1 and H_2 was motivated by our knowledge [54] that *if* the attractor is restricted to a manifold of dimension Q *and* the parameterization is evaluated only for states on the attractor, *then* perfect parameterizations of the form H_1 and H_2 (almost certainly) exist if $2Q < m$ or $Q < m$, respectively.

ι-shadow indefinitely; as we get more data we will find (almost certainly) that the probability of the data given the model goes to zero; not just for these particular models, but for every model in the model class(es) under consideration. Although it is not clear if there is a natural definition of the best model, at least on time scales much less than the recurrence time, it does seem likely that an ensemble over models will out-perform the best model for most reasonable definitions of "best."

And there is no simple stochastic fix. While adding a random component to a deterministic model may imply that a trajectory which stays near the verification exists, such a trajectory cannot be said to ι-shadow unless the random innovations required are consistent with the source of stochasticity specified by the model. For the AR parameterization above, the innovations must be consistent (at some confidence level) with I.I.D. drawn from an *a priori* specified Gaussian distribution. When model trajectories are restricted to remain on an unspecified manifold, the construction of stochastic terms which respect this constraint appears nontrivial. In practice, the stochastic models we have explored in this context are consistently over-dispersive in model-state space.

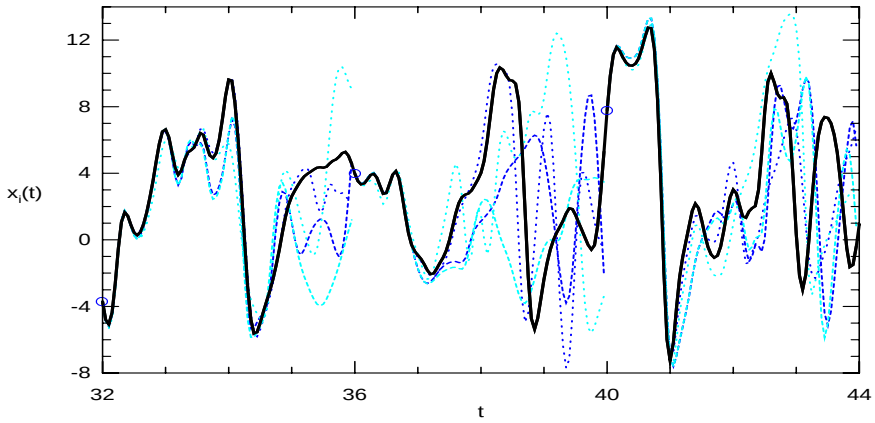

FIGURE 2.8. The trajectory of Figure 2, this time showing truth (solid) and 3 forecasts of 4 imperfect models: Linear (dark dotted), Constant (light dotted), IID (dark dashed), and AR1 (light dashed). At each of the circles an ensemble of model trajectories is initiated using the exact values of $\tilde{x}$, one trajectory for each model .

Forecast ensembles over models are shown in Figure 2.8; here there is no observational uncertainty when the model is initialized: each initial condition is exact (*i.e.,* the analysis corresponds to the true state projected into the model's state space). Since each of the individual models is wrong, the PDF will be incorrect, furthermore there need be no initial condition for any imperfect model which will ι-shadow for the duration of the forecast. And finally, no analog of the perfect ensemble exists; since an arbitrarily good weather model can have a horrid climatology. No model ensemble scheme will verify accountably, nor can a level of accuracy can be deter-

mined *a priori* which will guarantee a bound on the uncertainty of the forecast at a fixed future time. The only option for an accountable ensemble is to wait for physical analogs, as suggested by Lorenz [35] in terms of a single forecast; for daily weather, that may take some time.

Of course, a forecast trajectory in the future need not be held to the same standards as analysis trajectory of the past. While an ι-shadowing trajectory must exist for coherent variational assimilation, forecast models unable to produce an ι-shadow trajectory may still be of great value. A less constrained notion of shadowing is required: ϕ-shadowing.

2.5.1 ϕ-shadowing

A model ϕ-shadows for a time τ_ϕ if the model contains an initial condition, consistent with the initial observational uncertainty, which resembles the future closely enough for a forecaster: a ϕ-shadow need only be useful. This differs from ι-shadowing, for example, in that relatively large errors in the time of onset of a rain storm may be accepted (as long as a storm is forecast), or relatively small variations in the strength and location of spatially coherent structures may be acceptable (although these errors often result in huge contributions to the RMS error of a predicted field). The idea is to adopt an operationally relevant definition of what constitutes a useful forecast trajectory (see Murphy [44] for a discussion of what constitute a good forecast); determining the median of the distribution of times $\tau_\phi(\tilde{\mathbf{x}})$ over which a given model can ϕ-shadow provides a bound on predictability beyond which the use of Monte Carlo ensembles is, at best, questionable. For it is difficult to conceive of a useful purpose for ensemble forecasts beyond the time horizon over which the model can ϕ-shadow: if there is no initial condition which will reflect the future even roughly, what can be gained from a distribution of such model error dominated trajectories? This only argues for the existence of a ϕ-shadow, ideally the higher the probability of finding them, the more useful the model.

Requiring a ϕ-shadowing trajectory to exist is a much looser constraint than requiring a model to ι-shadow. A model ι-shadows for a time τ_ι if it contains a trajectory which is consistent with the observational uncertainty at *all* times t, $0 \leq t \leq \tau_\iota$. The tolerance is set by the uncertainty in the observations[9], and will be much more restrictive than just requiring a useful forecast [20, 56, 58]. If the observations are limited only by quantization uncertainty due to truncation, then to ι-shadow a trajectory must fall into

[9]In real applications, of course, the real measurements are rarely equivalent to the variables in the model-state space; there is an entire field of endeavor dedicated to relating point-wise physical measurements to grid point model variables, and the relation of both to the three-dimensional fields they are taken to represent. On which natural length-scale can one coherently define wind? or identify it with the velocity variables evolved within a model?

the quantization box corresponding to the observation at each observation time. For Gaussian uncertainties, some confidence level α must be chosen. Experience indicates that shadowing time is not very sensitive to this choice; at least in low-dimensional systems, once things go wrong, they go badly wrong rather quickly[10]. Given two diffeomorphisms, Anosov [8] and Bowen [10] determine sufficient conditions to guarantee the existence of another type of shadowing, but this ϵ-shadowing contrasts the trajectories of two well-defined mathematical systems; it is based upon assumptions which make it irrelevant (although nevertheless comforting) when contrasting imperfect models and real data (see Gilmour [20] for additional discussion). In practice, what we require are more vague but still quantifiable shadows, more along the lines of Eddington's use of the term [15]. Finding shadows is not so enlightening as realizing when they do not exist.

2.5.2 Bounding Boxes

Examining worst case scenarios is another common goal in weather forecasting. One approach would be to use the ensemble to define a region of model-state space within which the future is likely to fall[11]. The obvious method (see Figure 2.9) is to construct a convex hull from the ensemble members, and this is very useful in low dimensional models. But inasmuch as it requires $m + 1$ points to define a convex hull in m dimensions, this approach is untenable with ensembles of about 10^2 members in a typical model where $m \approx 10^7$.

An alternative to the convex hull is the bounding box, which has the advantage that it requires only two points and a coordinate system, regardless of the dimension of the model-state space. Consider a model on a spatial grid; at each grid point (for each variable) take the maximum value over all ensemble members: this co-dimension one plane is one "side" of the box, while the plane corresponding to minimum value forms the side opposite. Repeating this for all variables defines a volume of state space. Figure 2.9 uses the ensemble of Figure 2.6 to show both the convex hull (solid boundary) and bounding box (dotted boundary) defined by this ensemble. Note that the verification (which is truth, in this case) falls just outside the bounding box, but since the ensemble members were drawn from the same distribution as truth, then truth is no more likely to fall outside than any member of the ensemble. In cases where both are defined, the bounding box always has a greater volume than the convex hull, of course, and hence provides a larger (*i.e.,* easier) net with which to bag the

[10]Note, however, that the sci-fi models of Judd and Mees [27] in this volume address this problem explicitly.

[11]More precisely, to define a region which will contain the verification at the α level of confidence; ideally α is 100% but finite ensembles make this unlikely even in the perfect model scenario.

verification. We have not found this advantage to result in over-confidence in the model when the bounding box test is applied to forecasts of real data.

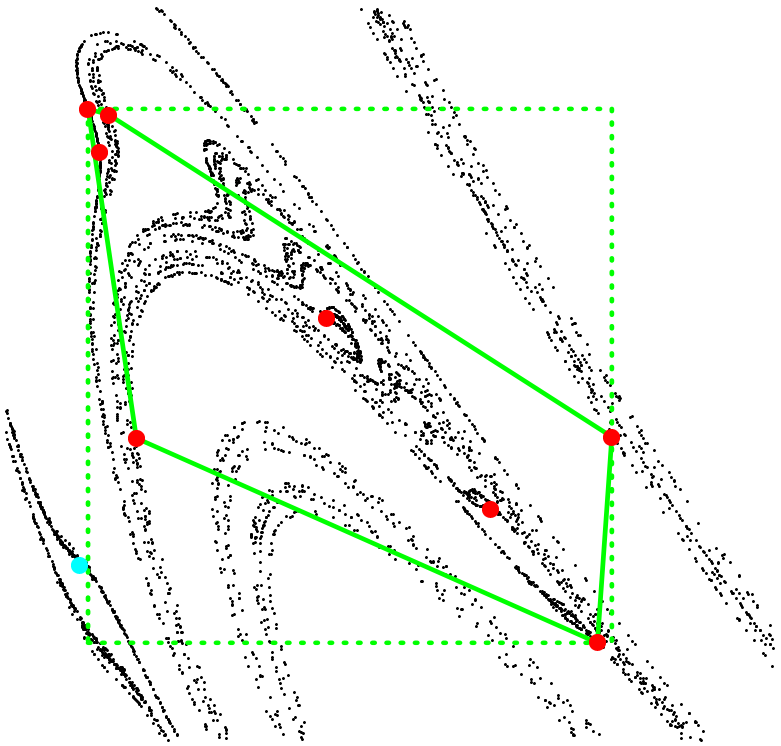

FIGURE 2.9. The convex hull (solid boundary) and bounding box (dotted boundary) for the ensemble of Figure 2.6.

In this scenario, it is straightforward to estimate the number of ensemble members required to have, say, a 95% chance that truth falls within the bounding box defined by the ensemble. This can be done analytically when the distributions are Gaussian as a function of standard deviation and bias [29]. Indeed, we plan to use this result to estimate bias of operational NWP ensembles. Of course, adopting the spatial grid carries the added bonus that so long as the verification is within the bounding box of the ensemble forecast, then nothing unexpected can happen. The target here contains much less information than an accurate PDF forecast, but in the imperfect model scenario there is no accurate PDF to be had.

2.5.3 Applications to Climate

Bounding boxes may also be of use in climate modeling. Typically one does not expect to find a ϕ-shadow in the climate model context where the goal is to reproduce the general statistics of likely states rather than a particular forecast trajectory. Lorenz [36] refers to this goal of climate modeling as predictions of the second kind. Yet the atmospheres of climate models may still be quite sensitive to initial condition, even when forced by observed sea surface temperatures (SST). Further, climate models are often run in an ensemble mode over historical periods, say, an ensemble of fifty year runs where each member is started from the analysis corresponding to a different day in 1950. While one should not expect to run large enough ensembles to produce even a ϕ-shadow over a 50 year period, it is reasonable to ask how large an ensemble would be required so that the analysis (or reanalysis [30]) for January 15, 1970 falls within the bounding box of all climate states of all ensemble members taken on all days between, say, January 1 to January 31, 1970. The point being that if reality (or even the analysis) consistently falls outside this bounding box, then the (dynamical) statistics of the climate model would be placed in doubt. Identifying specific historical periods where the model consistently fell outside the bounding box might aid in the identification of physical processes (active during those periods) which were insufficiently reproduced in the model. Over the historical record, one might hope for return of skill in the climate ensemble, inasmuch as each member is guided by the observed SSTs; in a free running fully coupled model the minimum size of the ensemble required to obtain a bounding bounding box would again be of interest in estimating the additional length of time (or number of ensemble members) that would have to be run to explore the additional degrees of freedom released.

Both the MST and the bounding box can be used to investigate natural variability of the climate system, either over time or in establishing whether the January 1 anomalies, variance adjusted, are the same in distribution as those of August 1. Given the short duration of many climate records, it is not uncommon to combine data from different seasons, once each data point has been adjusted to "remove seasonality." Examining the relative frequency with which the data from one calendar day fall into the bounding box defined by data from the day 6 months later, or the MST equivalent, would provide a useful check on whether simply adjusting the mean and variance is sufficient.

2.6 Multi-model CPT Ensembles

In this section a new method of truly multi-model ensemble forecasting is presented which attempts to take the limitations discussed above seriously. If we accept that each of our models is incorrect, that the "correct initial

PDF" is as ill-defined as the "true initial state", then we can construct a multi-model forecasting scheme which will outperform any individual model both in terms of ϕ-shadowing and in terms of the duration for which the verification remains within the bounding box defined by the ensemble members, at least in the limit of huge ensembles.

The simplest reaction to having M models is to identify the best one, discard the others, and compute $M \times N$ member ensembles under this single "best" model. If the models are of comparable quality, then it is likely that different models will tend to do better in different regions of state space (*i.e.,* on different days), due to variations in the particular processes that are important locally. In practice, there is rarely enough data to identify which one will be the best on a given day, and a reasonable alternative is to compute M, N-member ensembles, one ensemble under each model. Note that neither approach can produce a ϕ-shadow longer than the longest ϕ-shadow found within the individual models. If the M models really do have independent shortcomings (ideally, if they fail to ι-shadow in different regions of state space), then it is possible to cross-pollinate trajectories between models in order to obtain truly multi-model trajectories that explore important regions of state space the individual models just can't reach. This Cross-Pollination in Time (CPT) approach can outperform both of the methods above.

The basic CPT approach first takes the M N-member ensemble forecasts made under each model and combines them to form one large set of $N \times M$ points in the model-state space. This large ensemble is then pruned back to N member states, attempting to maintain a large bounding box while deleting one member in each pair of relatively close ensemble members (the details of the PDF are wrong anyway). These N conditions are then propagated forward under each of the M models. And so on.

Inasmuch as the CPT ensemble model implicitly contains all trajectories of each of its constituent models, CPT can ϕ-shadow as long or longer than any of the individual models. Similarly, in the limit of large ensembles, or in the absence of pruning, a CPT bounding box will contain the bounding box of the best model; and it is expected that given a good pruning scheme, the bounding box of the CPT ensemble will be more likely to contain the verification over a longer duration than those of an $M \times N$ member ensemble under the "best" model. While the optimal pruning scheme is still an object of research, the simple approach of taking the nearest pair of points, and then deleting the one of these two points with the smallest second nearest neighbor distance, has been found to work fairly well in some simple examples. Note that the aim of pruning is quite different than that of resampling from an estimated PDF [9].

This approach assumes that either all the models share the same model-state space, or the one-to-one maps exist which link their individual state spaces; neither needs be the case in weather forecasting. And, of course, when parameterizations of physical processes are involved one must con-

sider the time scale of pruning: for example, we would wish not to switch between parameterizations in the midst of growing a cloud. But at least the preliminary step of being able to run a collection of operational models on the same computer system has been achieved at the European Centre for Medium-range Weather Forecasting (ECMWF). Switching between models is a nontrivial process, and some may find it objectionable in principle as there is no longer a "set of equations" which are being solved, although one might argue that the process is, in fact, solving a rather large iterated function system. Ideally, future research will resolve the issues above while retaining a closed form for the model, if not the solutions. But it is interesting to note that physics of late has gotten along rather well without always building the mechanical model Lord Kelvin [33] held to be the prerequisite for understanding one hundred years ago. Perchance the twentieth century will be remembered as the "Century of The Equation."

2.7 Discussion

Chaotic systems are often thought unpredictable since they have the property of on-average exponential growth of infinitesimal uncertainties, at least when the (geometric) average is taken over a trajectory which explores the entire attractor. Yet this "exponential on average" growth places no bounds on (i) the growth of a finite uncertainty, or (ii) predictability over any finite time horizon, or (iii) the average uncertainty doubling time. The dynamics of uncertainty are much richer than simple uniform exponential growth [47, 60]. Lyapunov exponents are only *effective* rates, nothing need actually grow like $\epsilon_0 e^{\Lambda t}$.

Of course simple mathematical models of chaos, designed with tractability in mind, tend to have fairly uniform growth rates (by construction); this can yield a very biased picture of predictability. In the Baker's Map [49] for example, the fastest uncertainty doubling time of each initial condition is one iteration and the Lyapunov exponent is equal to one bit per iteration. Yet within the family of Baker's Apprentice Maps [59], all of which have a Lyapunov exponent greater than one, there are maps with arbitrarily large average doubling times [57]. *Even in* the Lorenz 1963 model, there are regions within which all perturbations must shrink for a (finite) time [60, 46]. Vannitsem and Nicolis [67] investigate these inhomogeneities in an atmospheric model.

Only in systems where the dynamics linearized about a trajectory accurately reflect the true nonlinear dynamics at macroscopic scales of interest, do Lyapunov exponents have any impact on predictability. As long as uncertainty stays infinitesimal, it cannot limit predictability, and once it is finite the Lyapunov exponents need not provide a reliable guide for uncertainty growth. Whether locally defined [3, 70] or global [14], Lyapunov exponents

are only *effective* rates, and even when infinitesimal perturbations really do grow exponentially in time, the uncertainty growth may saturate at an amplitude much less than the diameter of the attractor. This is clear from the macroscopic structure visible in Figure 2.4. In general, there need be no "Lyapunov Horizon".

Orrell contrasts forecast uncertainty growth due to chaos with that due to model error [48]. Assume for the moment that the model is perfect and an that forecast error does grow as

$$\langle \epsilon(t) \rangle = \epsilon_0 e^{\Lambda t}. \tag{2.5}$$

In this case, the value of $\langle \epsilon(t) \rangle$ at any fixed t can be made small by taking a sufficiently small ϵ_0. If the model is not perfect, however, there will be a difference between the velocity of the model trajectory in the model-state space and that of the system trajectory (projected into the model-state space). This difference remains even when $\epsilon_0 = 0$ leading to an initial error growth which is linear in time thus (initially) greater than $\epsilon_0 e^{\Lambda t}$. Thus for an imperfect model

$$\langle \epsilon(t) \rangle = \bar{v} t \tag{2.6}$$

where the value $\bar{v}$ is the magnitude of this velocity difference averaged over the projection of the system's invariant measure into the model-state space[12]. For an imperfect model, $\bar{v} > 0$ and therefore the forecast uncertainty due to model error will always dominate the forecast uncertainty due to chaos for sufficiently small ϵ_0. While "chaos" can make the error growth greater still, as $\epsilon_0 \to 0$ model error will dominate. Worse, it is not clear how to correct this with ensemble forecasts.

Given a perfect model, one might construct a perfect ensemble; but even if the model structure is correct and only the values of model parameters are uncertain, accurate PDF forecasts seem beyond reach. One may sample the parameter space in a sensible way, and construct a perfect ensemble for each realization, but the resulting ensemble PDF will not accurately reflect the likelihood of finding the properties of the future trajectory which will be observed. It is not obvious how to construct ensembles over model class.

Is the model class of deterministic systems too small? Perhaps [64], but it is not clear how to best introduce stochastic dynamics in structures where a strong deterministic nonlinear component is easily extracted; this is particularly the case when the deterministic dynamics are known to lie on a lower dimensional manifold, the details of which are not known. Other contributions in this volume [18, 19, 69] suggest a number of avenues. The operative question is how to best model the phenomenon: the issue of whether a real system "really is" deterministic or stochastic cannot be resolved from real data [39, 57]. Determining how to best model a phenomenon turns on the issue of how we decide to evaluate our models. This paper is intended to

[12] Orrell [48] illustrates this relationship in the Lorenz 1996 system, deriving the variation in $\bar{v}$ as a function of the parameter F in equation 2.1.

stimulate debate on sane methods of model evaluation; there may be no best.

Is an accountable probability forecast a viable goal? Perhaps not. The rank histogram evaluation techniques of Section 2.4 assume that "truth" is indistinguishable from the members of the ensemble, that all are drawn from the same distribution. This is never the case in practice, where we begin with uncertainty over initial condition, boundary condition, parameter, and even model structure. The verification is never "just another member" drawn from this distribution. For perhaps the one thing we are certain of is that our model class is incorrect: the very structure of our models will change with additional observations. This eventuality need not stop us from decreasing our uncertainty and refining our probability forecasts, but it will prevent our forecast PDF from producing flat rank histograms.

2.8 Summary

Chaos poses no difficulties for LaPlace's demon [31], whose abilities were such that given one exact snapshot of a dynamical system, a perfect forecast of the future could be calculated. Such a forecast is beyond the powers of a modern incarnation with the same abilities but without access to exact observations; even given imperfect measurements which stretch back into the distant past, she cannot determine the current state of the system from among a set of indistinguishable states. She can, however, foresee the probability of any eventuality. For mortals with imperfect models, even the foresight of exact relative probabilities is lost; we must expect to be surprised, occasionally, as there will be events which cannot even be foreshadowed.

As has long been recognized, uncertainty in the initial condition limits the utility of single deterministic forecasts of nonlinear systems like the Earth's atmosphere. If this uncertainty is accepted, then internal consistency requires that an ensemble of initial conditions, each consistent with the observations, be evolved forward under the model. Methods for selecting these initial conditions [38] have been advanced by Lorenz in 1965 and competing operational approaches dating back to the early 90's are used in the European and American weather forecasting centers. Assuming that the model physics is perfect, these methods aim at a weighted selection of the perfect ensemble [56, 57], where the weighting scheme depends on the aim of the forecaster.

Even under ideal conditions, uncertainty in the initial condition also limits the utility of single deterministic predictions of deterministic nonlinear systems; in practice ensembles of initial conditions are forecast with the dual aims of (1) estimating the reliability of that forecast and (2) estimating some aspects of the probability density function (PDF). Current rank histogram verification techniques are limited to scalar forecasts; Section 2.4.1 introduced a method using minimum spanning trees to allow computationally efficient verification in higher dimensional spaces, including the 10^7 dimensional weather model forecasts. Given a perfect model,

one may construct an accountable ensemble forecast system by sampling from a perfect ensemble; this scheme can yield accurate probability density estimates. In general, no perfect ensembles exist for imperfect models. If accountable estimates of the forecast PDF are unobtainable, we should question whether current skill scores provide a reliable guide for model improvement.

Ensemble prediction systems can consider Monte Carlo ensembles over initial conditions, parameters and model structure. If accurate probability forecasts prove untenable, what viable aims exist? Two have been discussed in Section 2.5: obtaining at least one good forecast trajectory (a ϕ-shadow), and constructing an ensemble whose bounding box is likely to contain the verification. In the long term, the bounding box of a large ensemble will evolve toward that of the climatology, containing all the observations and hence almost certainly containing whatever it is we are attempting to forecast; ideally we wish the box to grow as slowly as possible, but no slower.

Several shades of shadowing trajectory have been distinguished, and each has application in operational forecasting. The distribution of ι-shadowing times reflects the longest time scales over which there exists a model trajectory consistent with the observational uncertainties. How long can operational weather models ι-shadow? Inasmuch as variational data assimilation assumes ι-shadows exist, and may degrade the analysis if there is no ι-shadow over the entire assimilation window, knowledge of these time scales is of operational value, since ι-shadowing times would reveal limits to variational assimilation. ϕ-shadows need not stay so near the verification; indeed some practitioners at ECMWF already look for something similar to a ϕ-shadow when evaluating operational forecasts (Tim Palmer, personal communication). Their real value may come from examining historical data: if due to model error no useful forecast exists beyond some time scale, then what can model forecasts (ensemble or otherwise) possibly tell us regarding times beyond that horizon? This predictability horizon, the time scales at which the contribution of model error to the forecast is large compared to the natural variability of the system, is quite independent of time-scales derived from Lyapunov exponents. Sometimes greater, sometimes not. But the question is no longer the classic issue of not being able to find the correct initial condition; it is now an issue of there being no correct initial condition to be found.

Accepting the fact that an accurate PDF cannot be obtained allows consideration of other methods of evaluation. Two options are to examine the distribution of ϕ-shadowing times of each model, and to estimate the ensemble sizes required to obtain an ensemble bounding box which contains the verification at various lead times. A somewhat more drastic result follows from accepting the ensemble paradigm completely and considering not only ensembles over trajectories from different models, but even individual trajectories which are evaluated using multiple models, the CPT approach introduced in Section 2.6 being a naive first step in this direction. Nevertheless, CPT multi-model ensembles can outperform any individual model in terms of both ϕ-shadowing and producing a good bounding box, while unashamedly producing an ensemble mean that does not resemble the verification, and an MST rank histogram that is inconsistent with an

accurate probability forecast. Accepting the limits which exist even in ideal scenarios will force us to reevaluate the aims and evaluation of operational forecasting. Failure to do so is madness: there is no sane approach to an ill-posed goal other than to alter the object of the exercise.

Acknowledgements

It is a pleasure to acknowledge many enlightening discussions with over participants at the Newton Institute over the duration of the program. I am particularly indebted to I. Gilmour, J. Hansen, A. Hero, K. Judd, B. Malamud, A. Mees, R. Smith, and P. Young. My understanding of the means and ends of operational NWP ensembles has been much improved though enjoyable discussions with T. Palmer and Z. Toth and by the epistemological worries of M. Allen. D. Orrell provided computations on the two-level Lorenz system and its models, and new ideas for estimating model error. I am also grateful for the shared insights of M. Berliner, D. Broomhead, R. Buizza, T. Hamill, P. McSharry, A. Provenzale and C. Ziehmann, and many participants who contributed to the High-Resolution/Low-Order (the tastes great/less filling) debate. I remain grateful to C. Sparrow for introducing me to Cambridge in the first instance. This work was supported by Pembroke College, Oxford and the ONR Predictability DRI under grant number N00014-99-1-0056.

References

[1] H. Abarbanel. Challenges in modelling nonlinear time series. In this volume.

[2] H.D.I. Abarbanel, R. Brown, and J.B. Kadtke. Prediction in chaotic nonlinear-systems: Methods for time-series with broad-band Fourier spectra. *Physical Review A*, 41(4):1782–1807, February 1990.

[3] H.D.I. Abarbanel, R. Brown, and M. B. Kennel. Local Lyapunov exponents computed from observed data. *Journal of Nonlinear Science*, 2(3):343–365, 1992.

[4] R. Ahuja, T. Magnanti, and J. Orlin. *Network Flows*. Prentice Hall, Upper Saddle River, NJ, 1993.

[5] J. L. Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9:1518–1530, 1996.

[6] J. L. Anderson. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Monthly Weather Review*, 125(11):2969–2983, 1997.

[7] J. L. Anderson and W. F. Stern. Evaluating the potential predictive utility of ensemble forecasts. *Journal of Climate*, 9(2):260–269, 1996.

[8] D.V. Anosov. Geodesic flows and closed Riemannian manifolds with negative curvature. *Proc. Steklov Inst. Math.*, 90, 1967.

[9] C. H. Bishop, B.J. Etherton, and S.J. Majumdar. Adaptive sampling with the ensemble transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, 1999. in review.

[10] R. Bowen. ω-limit sets for axiom A diffeomorphisms. *J. Diff. Eqns.*, 18:333–339, 1975.

[11] C. Brankovicand and T.N. Palmer. Estimates of seasonal predictability and predictive skill from ecmwf provost ensemble integrations. *Q. J. Royal Meteorol. Soc.*, 1999. in press.

[12] L. Brillouin. *Scientific Uncertainty and Information.* Academic Press, New York, 1964.

[13] M. Casdagli, S. Eubank, J.D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. *Physica D*, 51:52–98, 1991.

[14] J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617–656, 1985.

[15] A. Eddington. *The Nature of the Physical World.* J. M. Dent and Sons, London, 1935. Everyman's Library, Vol 922.

[16] E. S. Epstein. Stochastic dynamic prediction. *Tellus*, XXI(6):739–759, 1969.

[17] S. Eubank and J.D. Farmer. An introduction to chaos and randomness. In E. Jen, editor, *Lectures in Complex Systems*, volume Lecture II of *SFI Studies in the Sciences of Complexity.* Addison-Wesley, 1990.

[18] W. Fitzgerald. An introduction to Monte Carlo methods for Bayesian data analysis. In this volume.

[19] G. Froyland. Extracting dynamical behaviour via Markov models. In this volume.

[20] I. Gilmour. *Nonlinear model evalution: ι-shadowing, probabilistic prediction and weather forecasting.* D. Phil. Thesis, Oxford University, 1998.

[21] T. Hamill and S. J. Colucci. Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, 125:1312–1327, 1997.

[22] T. Hamill and C. Snyder. A hybrid ensemble Kalman filter/3D-Variational analysis scheme. *Mon. Wea. Rev.*, 1999. In review (5 Oct 1999), 43 pages.

[23] T. Hamill, C. Snyder, and R. Morris. A comparison of probabilistic forecasts. *Mon. Wea. Rev.*, 1999. In review (24 Feb 1999).

[24] J. A. Hansen. *Adaptive Observations in Spatially-extended, Nonlinear Dynamical Systems.* D. Phil. Thesis, Oxford University, 1998.

[25] J. A. Hansen and L. A. Smith. The role of operatoinal constraints in selecting supplementary observations. *J. Atmos. Sci.*, 1999. in press.

[26] P.L. Houtekamer, L. Lefaivre, J. Derome, H. Ritchie, and H. Mitchell. A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124(6):1225–1242, 1996.

[27] K. Judd, M. Small, and A. I. Mees. Achieving good nonlinear models. In this volume.

[28] K. Judd and L. A. Smith. Indistinguishable states I: Perfect model scenario. *Physica D*, 2000. in Review (3 Jan 2000), 21 pages.

[29] K. Judd and L. A. Smith. Towards forecasting bounding boxes: Applications to both weather and climate. 2000. In preparation for *J. Atmos. Sci.*, 8 pages.

[30] R. Kistler, E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, H. van den Dool, R. Jenneand, and M. Fiorino. 2000: The NCEP/NCAR 50-year reanalysis. *Bull. Amer. Meteor. Soc.*, 1999. In press.

[31] Marquis de Laplace, Pierre-Simon. *Théorie Analytique des Probabilitiés.* Paris, 1820. Reproduced in the Oeuvres complètes de Laplace, Paris, Volume 11, 1886.

[32] C. E. Leith. Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102(6):409–418, 1974.

[33] William Thomson Lord Kelvin ˙ *Baltimore Lectures on Molecular Dynamics and the Wave Theory of Light.* Cambridge University Press, Cambridge, England, 1904.

[34] E. Lorenz and K. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, 55:399–414, 1998.

[35] E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20:130–141, 1963.

[36] E. N. Lorenz. Climate predictability. In *GARP Publication No. 16*, pages 132–136. WMO, 1975. Appendix 2.1.

[37] E. N. Lorenz. Predictability - A problem partly solved. In *Predictability*. ECMWF, Seminar Proceedings, Shinfield Park, Reading, RG2 9AX, UK, 1995.

[38] E.N. Lorenz. A study of the predictability of a 28-variable atmospheric model. *Tellus*, XVII, 3:321–333, 1965.

[39] E. Mach. *Knowledge and Error*. Reidel, Boston, 1976. See page 208.

[40] C.J. Marzec and E.A. Spiegel. Ordinary differential equations with strange attractors. *SIAM J. Appl. Math.*, 38(3):403–421, 1980.

[41] P. McSharry and L. A. Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Phys. Rev. Lett.*, 1999. in press.

[42] H. Melville. *Moby Dick*. Oxford University Press, Oxford, 1998. Oxford World's Classics, (Opening quote from pg 189; see also pg 508).

[43] F. Molteni, R. Buizza, T.N. Palmer, and T. Petroliagis. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, 122:73–120, 1996.

[44] Allan H. Murphy. What is a "good" forecast? *Weather and Forecasting*, 8:281–293, 1993.

[45] Allan H. Murphy and R. L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338, 1987.

[46] J.M. Nese. Quantifying local predictability in phase space. *Physica D*, 35:237–250, 1989.

[47] J.S. Nicolis, G. Meyer-Kress, and G. Haubs. Non-uniform chaotic dynamics with implications to information processing. *Zeitschrift für Naturforschung*, 38 a:1157–1169, 1983.

[48] D. Orrell. *A shadow of a Doubt: Model Error, Uncertainty, and Shadowing in Nonlinear Dynamical Systems*. 1999. Transfer of Status Thesis, *University of Oxford*.

[49] E. Ott. *Chaos in dynamical systems*. CUP, Cambridge, 1993.

[50] T. N. Palmer, R. Buizza, F. Molteni, Y.-C Chen, and S. Corti. Singular vectors and the predictability of weather and climate. *Phil. Trans. R. Soc. Lond.*, A 348(1688):459–475, 1994.

[51] T.N. Palmer, C. Brankovicand, and D. Richardson. A probability and decision-model analysis of provost seasonal multi-model ensemble integrations. *Q. J. Royal Meteorol. Soc.*, 1999. in press.

[52] C. Pires, R. Vautard, and O. Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*, 48A:96–121, 1996.

[53] K. R. Popper. *The Open Universe*. Routledge, London, 1982. (Originally published in 1956).

[54] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65:579–616, 1991.

[55] L. A. Smith. Identification and prediction of low-dimensional dynamics. *Physica D*, 58:50–76, 1992.

[56] L. A. Smith. Accountability in ensemble prediction. In *Predictabilty*, volume 1 of *ECMWF Workshop Proceedings*, pages 351–368, Shinfield Park, Reading, UK, 1996. ECMWF.

[57] L. A. Smith. The maintenance of uncertainty. In G. Cini Castagnoli and A. Provenzale, editors, *Past and Present Variability in the Solar-Terrestial System: Measurement, Data Analysis and Theoretical Models*, volume CXXXIII of *International School of Physics "Enrico Fermi"*, pages 177–246, Bologna, 1997. Il Nuovo Cimento.

[58] L. A. Smith and I. Gilmour. Accountability and internal consistency in ensemble formation. In *Predictability*. ECMWF, Seminar Proceedings, Shinfield Park, Reading, RG2 9AX, UK, 1998.

[59] L.A. Smith. Local optimal prediction: Exploiting strangeness and the variation of sensitivity to initial condition. *Phil. Trans. Royal Soc. Lond. A*, 348(1688):371–381, 1994.

[60] L.A. Smith, C. Ziehmann, and K. Fraedrich. Uncertainty dynamics and predictability in chaotic systems. *Q. J. Royal Meteorol. Soc.*, 125:2855–2886, 1999.

[61] D.B. Stephenson. Correlation of spatial climate/weather maps and the advantages of using the mahalanobis metric in predictions. *Tellus*, 49 A(5):513–527, 1997.

[62] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the a djoint vorticity equation - part I. *Q.J.R. Meteorol. Soc.*, 113:1311–1328, 1987.

[63] P. D. Thompson. Uncertainty of initial state as a factor in the predictability of large-scale atmospheric flow patterns. *Tellus*, 9:275–295, 1957.

[64] H. Tong. *Non-Linear Time Series Analysis*. Oxford Univ. Press, Oxford, 1990.

[65] Z. Toth and E. Kalnay. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.*, 74(12):2317–2330, 1993.

[66] H. M. van den Dool. Searching for analogues, how long must we wait? *Tellus*, 46 A(3):314–324, 1994.

[67] S. Vannitsem and C. Nicolis. Lyapunov vectors and error growth patterns in a T21L3 quasigeostrophic model. *J. Atmos. Sci.*, 54:347–361, 1997.

[68] R. von Mises. *Probability Statistics and Truth*. George Allen and Unwin, London, 1957.

[69] P. Young. The identification and estimation of nonlinear stockastic systems. In this volume.

[70] C. Ziehmann, L.A. Smith, and J. Kurths. The bootstrap and Lyapunov exponents in deterministic chaos. *Physica*, D 126:49–59, 1999.

[71] C. Ziehmann-Schlumbohm. *Vorhersagestudien in chaotischen Systemen und in der Praxis - Anwendung von Methoden der nichtlinearen Systemanalyse*. PhD. Thesis, Freie Universität Berlin, 1994. Meteorologische Abhandlungen N.F. Serie A Monographien.