What might we learn from climate forecasts?

Leonard A. Smith*

Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, United Kingdom

Most climate models are large dynamical systems involving a million (or more) variables on big computers. Given that they are nonlinear and not perfect, what can we expect to learn from them about the earth's climate? How can we determine which aspects of their output might be useful and which are noise? And how should we distribute resources between making them "better," estimating variables of true social and economic interest, and quantifying how good they are at the moment? Just as "chaos" prevents accurate weather forecasts, so model error precludes accurate forecasts of the distributions that define climate, yielding uncertainty of the second kind. Can we estimate the uncertainty in our uncertainty estimates? These questions are discussed. Ultimately, all uncertainty is quantified within a given modeling paradigm; our forecasts need never reflect the uncertainty in a physical system.

"Laws, where they do apply, hold only *ceteris paribus.*"
Nancy Cartwright (1)

The traditional approach to climate modeling is to build the most complicated model that will fit inside the largest computer available, run it once, and see what happens. This approach yields a single "best-guess" forecast. Yet even in high school physics, we learn that an answer without "error bars" is no answer at all. Although it is a nontrivial task to assign relevant uncertainty estimates to imperfect models of chaotic systems undergoing transient changes in forcing, doing so is conceivable. One alternative to devoting all our resources to one best guess is to use the same computer resource to perform an ensemble of model runs. This alternative would, of course, require the use of simpler models, and a balance between running different initial conditions (to cope with chaos), different model parameterizations and parameter values (to identify tuning issues), and different model structures (to mitigate model error). A single best guess from a complicated model run without good uncertainty estimates is impotent, whereas a beautiful set of ensemble statistics on too simple a model is irrelevant. How do we go about assigning resource between these two extremes? And how can we tell which physical phenomena of economic and social interest our current models might be able to forecast?

At best, our models hold only in certain circumstances. This is true even for our "Laws of Physics:" Newton's Laws are still celebrated for their successful prediction of the planet Neptune, although two historical facts (that one of the scientists who predicted Neptune also predicted the planet Vulcan, and that Vulcan was "observed" for many years) are less commonly found in physics texts. In the case of Vulcan, the then known Laws of Physics were applied outside their range of validity. By its very nature, this kind of failure is inconceivable before it is observed to have happened; because we cannot assign a meaningful probability to this occurrence, all results at the boundaries of our understanding must be treated as fundamentally uncertain. To make any progress, we assume the rosy scenario holds: (*i*) nothing horrible happens that takes the model beyond its range of validity (e.g., no asteroid collides with the earth), and (*ii*) no small but crucial feedback mechanism is missing from our model (i.e., our model has a range of validity). As we are forced to assume the rosy scenario, we can never make objective probability statements on the basis of our climate simulations. What we can do is establish their internal consistency: we can determine for which phenomena and on which time scales our models might reflect reality.

Of course, climate questions of interest in economics and politics are usually posed with the aim of quantifying particular changes, say, in regional weather patterns, in the likelihood of extreme events, and so on. Even within the rosy scenario, we cannot hope to quantify changes in some phenomena of interest unless our model can capture (bound) those same phenomena in the historical record. Real statisticians will immediately object, of course, that capturing the phenomena "in-sample" does not guarantee our ability to capture it "out-of-sample," that is, in the future. This is true, but we are seeking only a necessary condition: if our models cannot capture the phenomena of interest over the data period from which the model was constructed, say 1950–2000, then those interested only in economic impacts should not even look at the statistics of those phenomena in 2000–2050.

What can the theory of nonlinear dynamical systems tell us about our models of the earth's climate? It can illustrate details of the complexity of this project by analogy. Although it is unreasonable to expect solutions to low-dimensional problems to generalize to a million dimensional spaces, so too it is unlikely that problems identified in the simplified models will vanish in operational models. In the next two sections, we first step through the issues involved for any nonlinear dynamical system and then introduce a second kind of uncertainty. The fourth section illustrates these words in symbols, introducing a particular system/model pair. Mathematicians are always at a disadvantage to physicists when studying model error (i.e., the difference between the real system and a particular model). Physicists can always resort to real data from a real system, in which case no perfect model exists and model error is unavoidable; mathematicians tend to create a nice mathematical system and then a family of models. Often the model and the system are in fact the same thing: this is the perfect model scenario (PMS). But even when the models are intentionally imperfect, as they are in the example below, they are imperfect in a very special way. For this reason, attempts to model very simple physical systems, like electronic circuits (2), can more closely resemble climate modeling than, say, contrasting the behavior of different climate models, each constructed by different scientists who all share a similar education. The penultimate section turns to the climate problem in particular and examines the options for applying these ideas. Indeed, some major climate centers are already running (small) ensembles, and the climateprediction.com (http://www.climateprediction.com) project aims to

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Self-Organized Complexity in the Physical, Biological, and Social Sciences," held March 23–24, 2001, at the Arnold and Mabel Beckman Center of the National Academies of Science and Engineering in Irvine, CA.

Abbreviation: PMS, perfect model scenario.

*E-mail: L.Smith@lse.ac.uk.

run a huge ensemble experiment. The final section provides a few conclusions and an overview.

Predicting Chaos

Consider the earth's climate system as a nonlinear dynamical system. The current state of the atmosphere, ocean, biosphere, and so on, is contained in a single vector $\tilde{\mathbf{S}}$. This vector denotes one point in a state space, so named because each point in this space completely specifies one state of the system. As time passes, the state evolves, and we have a trajectory $\tilde{\mathbf{S}}(t)$, where t represents time. This is a trajectory in state space. It is, of course, unlikely that such a deterministic state space exists, because it would have to include, among other things, us. Nevertheless, assume for the moment that the world is deterministic in the sense of Laplace (3); then there exist deterministic "Laws of Physics" such that given the starting point $\tilde{\mathbf{S}}(t_{now})$, the future trajectory is completely defined by some mathematical function $\tilde{\mathbf{F}}(\tilde{\mathbf{S}})$. The ultimate climate prediction, and the traditional aim of weather forecasting as well, is to compute $\tilde{\mathbf{S}}(t)$; from this trajectory, every climate statistic could be accurately estimated.

Of course, there is a snag: nonlinearity in general, and chaos in particular, would cause the forecast trajectory to go astray unless the initial condition $\tilde{\mathbf{S}}(t_{now})$ were known exactly. Even with a perfect model, a long series of uncertain past observations of $\tilde{\mathbf{S}}(t)$ will not define the future unambiguously (4). Thus, even if we knew $\tilde{\mathbf{F}}$ and had the computational ability to solve the equations exactly, the best we could hope for would be a probability forecast, which is, in fact, what operational weather centers attempt with ensemble prediction systems (5–7). There are at least two additional difficulties. The first, climate forecasts share with weather forecasts: model error (8, 9). This will explain the profusion of tildes. The second difficulty is posed by time-dependent forcing: external effects like the amount of CO_2 in the atmosphere. As most climate experiments aim to study transient behavior, any appeals to mathematical properties like ergodicity are misguided.

In the perfect model scenario, when $\tilde{\mathbf{F}}$ is known but $\tilde{\mathbf{S}}$ is not, traditional best-guess climate modeling makes one long computer run and then takes space and time averages in the hope that even though the model trajectory $\mathbf{S}$ is not close to the true trajectory $\tilde{\mathbf{S}}$ at each point in time, their averages are similar. Given only a single run, it is not clear how to determine just how much averaging is required. In any event, many statistics of economic interest simply cannot be computed from such averages: a perfect forecast of the monthly average temperature in Berlin simply cannot tell us the number of days on which construction was halted because it was too cold for cement to set. For many statistics of economic interest, the statistic evaluated either on a time average or on a climate mean state is meaningless. Uncertainty of the mean value says nothing about the likely variations observed day to day.

Alternatively, we can try to maintain our uncertainty by keeping track of how it evolves with time (7, 10). One approach is to keep track of this uncertainty analytically, but climate models are nonlinear: we cannot assume that the superposition of true trajectories is a true trajectory. Nor can we assume that small perturbations have small effects. Thus within a matter of days, at best (11), we must track the uncertainty manually. This may be done, for example, by computing an ensemble of trajectories, each consistent with all our observations. There are additional difficulties in forming a perfect ensemble (12), because the nonlinearities link the longest time scales to the shortest forecasts (13), but we will ignore these for the time being.

A major advantage of the ensemble approach is that it removes the *a priori* need to average over time—a real benefit in a transient forcing experiment. Given 100,000 ensemble trajectories, each started in the 1950s, we can examine the distribution of model temperatures of Glasgow at 7:30 p.m. on August 8, 2001. Given a coupled nonlinear simulation, it is not clear that one can get the averages right without being able to simulate the details correctly. To quantify how well a model is doing, one can define a temporal credibility ratio,

$$\tau_{\text{cred}} = \frac{\Delta t}{\tau_{\text{ave}}}, \qquad [1]$$

where Δt is the smallest time step in the model, and τ_{ave} is the smallest duration over which a variable has to be averaged before it compares favorably with observations. Most variables of economic interest will have different temporal credibility ratios depending on the spatial length scale of interest; extent and duration must be considered together. Given only a single climate model run under a transient forcing scenario, there are good reasons to argue that τ_{ave} must be fairly large on statistical grounds. Most of these reasons vanish, however, if an ensemble run is made. The simplest question is to ask whether an observation falls within the range of the ensemble members: at least one member above it and one below. Was the average August temperature over Europe in 2001 within the range of the ensemble values? If not, then what reason might we give for that particular averaged variable in 2050 to bear any relation to the eventual observation? In this way, we can begin to estimate the space and time scales on which our models have skill and thus their limitations as quantitative tools for policy. Note that the ability to bound does not require an accurate probability forecast, whereas the inability to bound can assist in directing model improvements.

Of course, we are far from having $\tilde{\mathbf{F}}$; denote this year's best state-of-the-art model as $\mathbf{F}$ and vectors in the corresponding model-state space by $\mathbf{S}$. At best, $\mathbf{S}$ is a projection of $\tilde{\mathbf{S}}$, and projections have rather nasty mathematical properties. For example, we cannot meaningfully follow a true probability density function in the projected space. Indeed, much confusion has come from the fact that our model variables have the same names as the physical variables: variables like "temperature" and "wind speed" mean very different things in a model where a grid point corresponds to 100-by-100 kilometer box.

Uncertainty of the Second Kind

Lorenz (14) distinguishes predictions of the first kind, where the time order of individual forecast trajectories is important, from those of the second kind where the goal is a probability distribution over the states of the system. To some extent, the two merge together when making ensemble forecasts under a perfect model.

The climatology is the distribution of all physically relevant states of the system in state space. This distribution is reflected by the system's attractor, if such a thing exists. In practice, the climatology is often approximated by the distribution of the historical observations. Traditionally, forecast errors have contrasted the forecast (either probabilistic or best guess) with random draws from the climatology. Uncertainty of the first kind addresses the question of uncertainty of the future state within some distribution of possible states like the climatology or an ensemble weather forecast. Uncertainty of the second kind considers not our uncertainty in the state of the system, but the uncertainty in the likely distribution of states, effectively our uncertainty in the climate of the distant future. Model error provides a major source of this uncertainty (8, 13).

Given a good physical model with only small uncertainty in the parameters, and knowing the initial condition of atmosphere and ocean for every Monday in the 1950s, what is the uncertainty in the distribution of climate variables in the year 2000? What most contributes to this uncertainty? Differences in the initial conditions? in the parameters? in the forcing? The light that

ensemble climate experiments can shed on these questions is discussed in the next section. It is important to remember that a best-guess experiment must assume that the trajectory is (*i*) realistic, and (*ii*) representative, and also that (*iii*) the future is rosy. Ensemble climate experiments relax point (*i*) and test (*ii*) but do not mitigate (*iii*).

There are many more ways to be wrong in a 10^6 dimensional space than there are ways to be right. But there is also a difference between not knowing where to start, and no such place existing. The first is a question of state estimation, the second a question of shadowing, and thus model error. Any initial state that yields a trajectory consistent with the available observations is said to ι-shadow (12, 13, 15) the true trajectory or, more correctly, the true trajectory projected into model state space. Within PMS, there will be a set of indistinguishable states that will shadow forever (4). Without PMS, the duration over which the observations can be shadowed, taken over all initial system states, provides a measure of model error (9, 15). Given the nonlinearities involved, it is not clear whether a model that cannot produce reasonable "weather" can produce reasonable climate statistics of the kind needed for policy making, much less whether it can mimic climate change realistically.

Difficulties of Modeling by Analogy

Obtaining a quantitative illustration of model error requires one either to consider a real physical system or to make up a mathematical system and then pretend to forget that a perfect model exists. The second approach is taken here by defining a simple nonlinear two-dimensional chaotic map, including a seasonal cycle as the system. A "first-principles" model with an imperfect but mathematically relevant structure is then created for this system. The challenges of using very good but imperfect models of both weather and climate change are then examined by analogy. In place of the real world system $\tilde{\mathbf{F}}$, take the equations

$$\tilde{x}_{i+1} = 1 - a\left(c\,\sin\left(\frac{\tilde{x}_i}{c}\right)\right)^2 + \tilde{y}_i + e\tilde{\nu}_i \qquad [2]$$

$$\tilde{y}_{i+1} = b_i c\,\sin\left(\frac{\tilde{x}_i}{c}\right), \qquad [3]$$

where the index i denotes time, and $\tilde{\nu}$ is a random variable that allows for external perturbations to the system. Note that it would be unrealistic to assume $\tilde{\nu}$ is an independently distributed standard normal random variable. Also recall that for sufficiently large c, $c\,\sin(x/c) \approx x$. To reflect a seasonal cycle, assume

$$b_i = \bar{b}\left(1 + d\,\cos\left(\frac{2\pi i}{12}\right)\right). \qquad [4]$$

Initially we will consider a purely deterministic system ($e = 0$), with parameter values $a = \bar{a}$, $\bar{a} = 1.35$, $\bar{b} = 0.2$, $c = 8$, and $d = 0.5$. Note that if $a = 1.4$, then in "perpetual January" mode (where the cosine term is replaced by unity), the system approaches the well studied Hénon map (16) as $c \to \infty$. The perpetual January attractor will, of course, differ from the January slice of the true attractor.

How do we build a model? First pick a model class, say two-dimensional polynomial maps with a third seasonal component. Some parameters will be estimated from the data, others may be known exactly within the context of the model (e.g., the number of months in a year). Our first model has the structure:

$$x_{i+1} = \phi - \alpha x_i^2 + \gamma y_i + \varepsilon\nu_i \qquad [5]$$

$$y_{i+1} = \beta_i x_i, \qquad [6]$$

where ν is an independently distributed standard normal random variable. Given that the system is deterministic (that is, $e = 0$), it might seem reasonable to set $\varepsilon = 0$, thereby making the model deterministic. Reasons for not doing so when forecasting will be presented elsewhere. The model's seasonal cycle is

$$\beta_i = \beta_0\left(1 + \delta\,\cos\left(\frac{2\pi i}{12}\right)\right). \qquad [7]$$

How do we determine the parameters or, even more interestingly: what are the "correct" values? As argued elsewhere, there are no correct parameter values given an imperfect model (13); the best values depend on the application of interest. This suggests an ensemble over parameter values, as well as over initial conditions. Note that the Bayesian agenda provides a nice framework for generating sets of parameters but recall that the probability of every set of parameter values given the data will be zero in the long run: there is no perfect model within the available model class, and there are no infinitely long shadows.

In this simple example, the state space of the system and that of the model both have dimension two, but x is fundamentally different from $\tilde{x}$ even if they have the same name. Among other things, the system and the model have different attractors. In general, attractors are defined by the dynamics of the system over rather long time scales, durations similar to the time it takes the system to return near the same point in state space. There is certainly no such recurrence in a climate change experiment, where some parameters change with time in a transient fashion. For the earth's atmosphere, the recurrence time is estimated to be longer than the lifetime of the planet [longer, in fact, than the expected lifetime of the universe (17)]. And nonlinearity can link the longest time scales to the shortest forecast horizon (12, 13). Long recurrence times in nonlinear systems make Occam's razor a rather blunt instrument for model building.

To extend the analogy to include climate change, let a become a function of time, specifically $a(t) = \bar{a}$ for $t < t_0$, whereas $a(t) = \bar{a}(1 + \bar{\Delta}t)$ for $t > t_0$, with the arbitrary choice $t_0 = 1950$. We adopt a similar form for $\alpha(t)$ and assume that we know the rate of increase exactly, that is, $\Delta = \bar{\Delta} = 0.1$ per century [or $0.1/(100*12)$]. We can now study climate change in this simple analog system and examine the challenges we are up against.

Suppose we have about 100 years of data from 1900 to 2001 that are used to determine relatively likely sets of parameter values. Given a model, we can run an ensemble of packages, each consisting of one parameter set, and an ensemble of initial conditions. Do runs under this model structure bound the observed monthly values of $\tilde{x}$ and $\tilde{y}$ [taking into account the effects of the finite size of the total ensemble, any observational error, and the need to use bounding boxes when assessing anything more than scalar variables (13)]? If not, then compute seasonal (or annual) averages and repeat the tests. Knowing which phenomena this model structure can (and cannot) bound in free running hindcast mode tells us a great deal regarding what it might bound in forecast mode. Examining the distribution of the ensembles tells us about the sensitivity of the model structure; it may also provide hints regarding the sensitivity of the system. This provides a lower bound on how much we should trust the results and at least carries us beyond the requirements of high school physics experiments.

What then is climate change? Given the perfect model scenario with both $e = 0$ and exact initial conditions, climate change is no more than the difference between two trajectories, one with constant forcing and one with transient forcing. When retaining perfect model scenario while allowing either $e \neq 0$ or some fundamental uncertainty in the initial condition [perhaps because of quantum mechanics, as suggested by Lorenz (14)], things begin to get interesting. In this case, there is a many-worlds interpretation of the ensemble consisting of all system trajecto-

ries consistent with $\tilde{F}$ given $\tilde{S}$ (t_{now}). Some interpretations of quantum mechanisms state that every measurement differentiates "us" into one of several distinct universes—which one depends on what has happened in our particular universe (18). As these universes are initially identical and share the same Laws and Physics, there is a subset of them that corresponds to the perfect model scenario ensemble above. Unfortunately, each "we" is unlikely to obtain information on more than one member of this many-worlds ensemble. It is interesting to note that Deutsch has a somewhat different notion of the magnitude of a small uncertainty than that held by most meteorologists, leading him to the conclusion that slightly different initial conditions would yield only slightly different multiverses (i.e., similar evolved perfect ensembles).

In the case of constant forcing, incomplete information regarding the initial condition will yield a growth in uncertainty, and the ensemble of initial states will spread out on "the attractor." The information content of a forecast lies in the difference between the ensemble and the attractor. Although it is difficult to define what constitutes a good forecast (19), this difference will almost certainly become too small to be of any utility "soon" (see, however, ref. 7). If the forecast ensemble has a finite number of members, then as $t \to \infty$, the two distributions become indistinguishable, at which point any forecast is useless (12). But we are interested in finite times. In the case of transient forcing, there is no attractor. A perfect ensemble will spread out with time, but it will not be attracted toward any fixed set of points, as the parameters of the system are undergoing a monotonic evolution.

Thus we are left with three distributions, all of which include one coordinate that reflects the annual cycle, and two of which change from one "January" to the next. Name the attractor of the system with constant ($a(t) = \bar{a}$) forcing $\mu_0(\bar{a})$, the distribution evolving from $\tilde{S}(t_{now})$ under constant forcing $\mu_1(\bar{a}, \tilde{S}(t_{now}), t - t_0)$, and the distribution evolving from $\tilde{S}(t_{now})$ under the transient (climate change) forcing $\mu_2(a(t), \tilde{S}(t_{now}), t - t_0)$. The discussion above can be summarized as saying that, as $t \to \infty$, $\mu_1 \to \mu_0$, and that beyond the range of operational weather forecasting the best forecast is, in fact μ_0. For a constant forcing, the distribution $\mu_1(t)$ is the climate, whereas the trajectory $\tilde{S}(t)$ is the weather, hence the adage, "Climate is what you expect; weather is what you get."

Climate change in this case is the difference between μ_2 and μ_1 as a function of time. For a given $a(t)$, reality will trace only one trajectory; the uncertainty between that trajectory and the corresponding distribution is uncertainty of the first kind. In contrast, the uncertainty as to which distribution is relevant (in the simplest case, μ_1 or μ_2) reflects uncertainty of the second kind. Within the perfect model scenario, uncertainty of the second kind arises only from uncertainty in the forcing $a(t)$ if the distribution of states at initial time is known.

Although climate change is defined as the difference between distributions, one can always look at particular statistics, say the mean value of the distribution in state space, $\langle \tilde{S} \rangle_{\mu_1}$, or the mean value of a scalar like $\tilde{x}^3$, depending on one's interests. The important point here is that, whereas one can define various "climate statistics," climate itself is always a distribution. By construction, the best-guess trajectory is at best one draw from this distribution and as such cannot reflect its width, much less the uncertainty in its structure.

So what is the climate mean in our simple case? Consider the mean value of $\tilde{x}$; taking the mean over $\mu_0(t)$ will yield a strictly periodic function, denoted by $\langle \tilde{x}(t) \rangle_{\mu_0}$, which has a period of 1 year (i.e., 12 months). Repeating the calculation for $\mu_1(t)$ will yield $\langle \tilde{x}(t) \rangle_{\mu_1}$, a function of time that initially differs from the periodic function obtained previously but (rapidly?) approaches it as time passes.

And for $\mu_2(t)$? Here the average of $\tilde{x}$ shows an aperiodic annual oscillation along with a systematic increase between 1950 and 2000. There are three crucial points in connection with this behavior:

i. It is misleading to say that this increase is "superimposed" on a periodic cycle, because the system is nonlinear; we have no recourse to the superposition of states.
ii. In general, $\langle \tilde{x}(t) \rangle$ is not a solution to the equations of motion: the evolution of the mean under the model is not the mean of the evolving climate. The mean is unlikely even to lie on the attractor that defines the climate!
iii. The average $\langle \tilde{x}(t) \rangle$ cannot be extracted from any best-guess run.

Before leaving the perfect model scenario to rejoin our modeling analogy of the real world, we discuss the utility of climate mean statistics even if they are known exactly. Assume that the transient forcing is realized: what does knowing the many-worlds mean tell us about what we will experience in the future? Not much, really. Although the trajectories of $\tilde{x}$ in each of the many worlds may be interesting, superimposing all of them yields a wide distribution of possibilities for every month in the future, which will, among other things, bound reality. Their many-worlds mean, on the other hand, will bear little resemblance to any of them and will not reflect their variability or their extreme values, two issues of immediate economic and social interest. Given the distribution $\mu_2(t)$, we can compute the 5 and 95% (or 0.1 and 99.9) bounds on any variable of interest. We can also look back in time and see how often $\tilde{x}$ falls outside of these bounds in the past. Within PMS, this will happen about 10% (or 0.2%) of the time; departure from this target for operational models yields insight and suggestions for model improvement.

Now back to the analogy with the real world. Again we can run our model in ensemble mode, considering a range of reasonable values for the parameters. This is done for each package of model structure, parameterizations, and parameter values, which in turn considers an ensemble of the initial conditions. The model-climate statistics will, of course, share the limitations of the true climate statistics while adding new shortcomings of their own. They cannot, for example, be expected to reproduce the accountable probability density function (PDF) statistics given by the many-worlds ensemble. It is not clear how to interpret operational distributions using multiple packages and imperfect model structures; this distribution will not resemble the PDF of the system. Empirically, we can determine how well the ensemble bounds the observations (that is, attempt to compute the 95th percentile), noting that this will require some kind of model output statistics (MOS) (20) to account for projection effects. Indeed, MOS is simply a special case of the projection operator discussed above. Projection effects must also be accounted for in the context of fingerprint methods for detecting patterns of climate change (21, 22). In general, there remains the open question of whether any projection operator exists under which current climate models could be said to shadow the historical record.

In the simple model above, the ensemble runs under the model fail to bound the single run under the system designated as truth much too often. In particular, the model ensembles miss the extreme values in the observations. This is true in-sample, that is, when examining the historical data used to determine the best parameter values for each package. Given these results, it is reasonable to assume that a first-principles modeler would see where the problem was, namely that large values of x gave systematically worse predictions, the modeler might then guess that the forecasts would be improved by adding a higher order term in x. This suggests changing the model structure by replacing Eq. **4** with

$$x_{i+1} = \phi - \alpha x_i^2 - \kappa x_i^4 + \gamma y_i + \varepsilon \nu_i \quad [8]$$

In many ways, this second model will be better than our first model. But it remains imperfect; indeed, any polynomial model with a finite number of terms will be imperfect. The issue is one of model class. Let our model class consist of all polynomial models with terms up to 10th order: there is still no perfect model for any of the functional forms at our disposal. There is no question that the model class containing our computer-based climate models does not contain a member corresponding to the real world, and philosophers are divided as to whether any model class exists that contains any real physical dynamical system with sustained dynamics, even the electric circuit.

We must always work in the rosy scenario, but, by testing our ensembles in sample, by adjusting our aims to bounding boxes rather the probability density functions, we can do much useful work nonetheless. For instance, we can investigate just how bad uncertainty of the second kind is by the contrasting distributions from various packages. When the experiment is repeated with ensembles over model structure, each package again yields its own distribution; uncertainty of the second kind is reflected in how much these distributions differ. The two extreme options are (*i*) that they are each rather peaked with little overlap or (*ii*) that they are rather similar. It seems we must aim for *ii*, at least when grouping packages over model structure. Our competing model structures must be so good that the details are irrelevant (within the rosy scenario). If the details matter, we need to devote resources to model development, hoping that the existing data contain enough information to improve the model or at least disqualify some packages. Alternatively, if the distributions are broadly similar, we might investigate larger ensembles with the aim of finding what distinguishes them and hence get a handle on model improvement.

There is one complication we have not yet mentioned: how to contend with unphysical packages. These are models with parameter sets that appear reasonable enough but produce physically unreasonable results, say, when run from 1950 to 2000. For example, the first initial condition tried in a given package may crash the computer before the simulation reaches the year 2000. Such a package cannot be dismissed out of hand on the basis of one 1950–2000 run. It might be the case that all packages with this model structure become unstable with the same probability P_{diverge} per unit time. In that case, those that happened to make it to 2000 are no more stable in 2000–2050 that those that did not. Thus multiple initial conditions for each package are required, both to estimate P_{diverge} and to examine structure of the distribution even if $P_{\text{diverge}} = 0$. And how might one detect and account for a more subtle "unphysical" trajectory in 2000–2050?

Of course, the system may also be unstable! This fact is usually ignored in the rosy scenario, because there is little hope of simulating a system in a region of state-space far from the observations.

Operational Climate Modeling

Allen (23) has proposed distributing a climate model (one package from a small collection) and an initial condition to every interested individual over the World Wide Web, allowing a single personal computer to compute a single unique trajectory over the period 1950–2050. There are questions of experimental design still to be resolved (24), yet climateprediction.com can begin to address many questions in the context of what were recently state-of-the-art models. Regardless of what we learn about the future, such an experiment will teach us a great deal by contrasting the distributions of large model ensembles with the observed climate over the past decades. It will make it possible to see which variables (and credibility ratios) can be captured in the data used to construct the models.

It will undoubtedly be argued that the model(s) chosen for climateprediction.com are too simple, and that showing a simple model fails for a particular variable does not imply a newer bigger better model will fail. This argument is as irrelevant as it is true. If model results are being used as forecasts and not only for pure research, then they are incomplete without a reliable estimate of the forecast uncertainty of every variable communicated to policy makers. Arguably, there is at present no good baseline for any climate modeling scenario that establishes which variables and time scales are reliable. Composing a baseline for any good model strengthens our confidence in its "newer bigger better" offspring. Therefore the climateprediction.com experiment can also provide a realistic baseline for confidence levels for more complicated models until such time as the reliability of those models can be quantified directly. A great deal of work has gone into testing and verifying the components that compose state-of-the-art models, yet in a very real sense, as coupled nonlinear models per se, they are not yet out of high school, at least not in terms of the questions their designers are asked to answer.

Conclusion

> "There is no more common error than to assume that, because prolonged and accurate mathematical calculations have been made, the application of the result to some fact of nature is absolutely certain."
>
> Alfred N. Whitehead (25)

The perfect model scenario is a useful but misleading fiction. And there is no simple stochastic fix. This does not imply that increasing resolution, improving parameterizations, introducing stochastic physics, and the like, will not significantly improve our current models but it does suggest that careful thought is required in quantifying exactly what we mean by "improve."

When extrapolating into the unknown, we wish both to use the most reliable model available and to have an idea of how reliable that model is. One argument against the ensemble experiments above is that only a model much simpler (faster) than a current state-of-the-art model can be used. But what faith do we have in today's most complicated models, other than the fact that in 10 years time, well within the range of the forecast being made, this same model will be condemned as too simple to be worthy of serious study?

We have shown how ensembles can be used in-sample to identify minimum scales for averaging and discussed the assignment of resources between making more realistic models and making relevant quantitative estimates of just how realistic they are. If we cannot obtain an accurate probability forecast, then model improvements should be aimed elsewhere, perhaps to obtaining a bounding box. Feedback for model development is important; model "improvements" aimed at reducing forecast errors actually due to uncertainty in the initial condition may have already made our models overly stable.

Models, when they do apply, will hold only in certain circumstances. Belief in extrapolation outside observed circumstances is largely a question of faith: we cannot know *a priori* whether we are discovering Neptunes or Vulcans. We may, however, be able to identify shortcomings of our model even within the known circumstances and thereby increase our understanding.

I am happy to acknowledge fruitful discussions with M. Roulston, D. Orrell, and C. Bishop, and to thank M. Allen, J. Hansen, T. Palmer, A. Thorpe, and Z. Toth for decreasing my uncertainty regarding large models. I am a Senior Research Fellow of Pembroke College, Oxford, U.K. This work was supported by ONR Grant N00014-99-1-0056.

1. Cartwright, N. (1999) *This Dappled World* (Cambridge Univ. Press, Cambridge, U.K.).
2. Smith, L. A. (2000) in *Proceedings of IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (AS-SPCC)*, ed. Haykin, S. (IEEE, Piscataway, NJ), pp. 129–134.
3. Earman, J. (1986) *A Primer on Determinism* (Reidel, Boston).
4. Judd, K. & Smith, L. A. (2001) *Physica D* **151,** 125–141.
5. Palmer, T. N. (2000) *Rep. Prog. Phys.* **63,** 71–116.
6. Toth, Z. & Kalnay, E. (1997) *Mon. Wea. Rev.* **125,** 3297–3319.
7. Smith, L. A., Ziehmann, C. & Fraedrich, K. (1999) *Quart. J. R. Meteorological Soc.* **125,** 2855–2886.
8. Orrell, D., Smith, L. A., Barkmeijer, J. & Palmer, T. (2001) *Model Error in Weather Forecasting, Nonlinear Processes in Geophysics*, in press.
9. Gilmour, I. (1998) Ph.D. thesis (Univ. of Oxford, Oxford).
10. Smith, L. A. (1997) in *Proceedings of the International School of Physics "Enrico Fermi,"* Course CXXXIII, ed. Cini, G. (Società Italiana di Fisica, Bologna, Italy), pp. 177–246.
11. Gilmour, I., Smith, L. A. & Buizza, R. (2001) *J. Atmos. Sci.*, **58,** 3525–3539.
12. Smith, L. A. (1996) in *Predictability* (ECMWF, Shinfield Park, Reading, PA), Vol. 1, pp. 351–368.
13. Smith, L. A. (2001) in *Nonlinear Dynamics and Statistics*, ed. Mees, A. I. (Birkhauser, Boston) pp. 31–64.
14. Lorenz, E. (1975) in *The Physical Basis of Climate Modeling* (World Meteorological Organization, Global Atmosphere Research Program, Washington, DC), Publication 16, 132–136.
15. Gilmour, I. & Smith, L. A. (1997) in *Applied Nonlinear Dynamics and Stochastic Systems Near the Millennium*, eds. Kadtke, J. B. & Bulsara, A. (AIP, New York) pp. 335–340.
16. Hénon, M. (1976) *Commun. Math. Phys.* **50,** 69–77.
17. van den Dool, H. M. (1994) *Tellus A* **46,** 314–324.
18. Deutsch, D. (1997) *The Fabric of Reality* (Penguin, London).
19. Murphy, A. H. (1993) *Weather Forecast.* **8,** 281–293.
20. Glahn, H. R. & Lowry, D. A. (1972) *J. Appl. Meteorol.* **11,** 1203–1211.
21. Levine, R. & Berliner, L. M. (1999) *J. Climate* **12,** 564–574.
22. Hasselman, K. (1976) *Tellus* **28,** 473–485.
23. Allen, M. (1999) *Nature (London)* **401,** 642.
24. Hansen, J. A., Allen, M., Stainforth, D., Heaps, A. & Stott, P. (2001) *World Res.*, Rev. 13.
25. Whitehead, A. N. (1953) *Alfred North Whitehead: An Anthology* (Macmillan, New York).