Model error in weather and climate forecasting

Myles Allen[*]**, Jamie Kettleborough**[†] **and David Stainforth**[*]

[‡]*Department of Physics, University of Oxford*
[†]*Space Science and Technology Department, Rutherford Appleton Laboratory*

"As if someone were to buy several copies of the morning newspaper to assure himself that what it said was true" *Ludwig Wittgenstein*

1 Introduction

We review various interpretations of the phrase "model error", choosing to focus here on how to quantify and minimise the cumulative effect of model "imperfections" that either have not been eliminated because of incomplete observations/understanding or cannot be eliminated because they are intrinsic to the model's structure. We will not provide a recipe for eliminating these imperfections, but rather some ideas on how to live with them because, no matter how clever model-developers, or fast super-computers, become, these imperfections will always be with us and represent the hardest source of uncertainty to quantify in a weather or climate forecast.

We spend a little time reviewing how uncertainty in the current state of the atmosphere/ocean system is used in the initialisation of ensemble weather or seasonal forecasting, because it might appear to provide a reasonable starting point for the treatment of model error. This turns out not to be the case, because of the fundamental problem of the lack of an objective metric or norm for model error: we have no way of measuring the distance between two models or model-versions in terms of their input parameters or structure in all but a trivial and irrelevant subset of cases. Hence there is no way of allowing for model error by sampling the space of "all possible models" in a representative way, because distance within this space is undefinable. If the only way of measuring model similarity is in terms of outputs, complications arise when we also wish to use these outputs to compare with observations in the initialisation of a forecast. These problems are particularly acute in seasonal or climate forecasting where the range of relevant observational datasets is extremely limited. Naive strategies run the risk of using observations twice and hence underestimating uncertainties by a significant margin.

A possible way forward is suggested by stepping back to remind ourselves of the basic epistemological status of a probabilistic forecast. Any forecast of a single or seldom-repeated event that is couched in probabilistic terms is fundamentally unfalsifiable. The only way of verifying or falsifying a probabilistic statement is through examining a sequence of forecasts of similar situations and assessing the forecast system's "hit rate", and we cannot compute a hit rate with a single verification point [33]. The objection that there are lots of different aspects of a climate forecast that we could verify against observations is a red-herring, because all these variables are intimately interlinked. If we produce an ensemble forecast that purports to span a nominal 5-95% "range of uncertainty," it is clearly absurd to expect 90% of the

[*]Correspondence to: Myles Allen, Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK. E-mail: myles.allen@physics.ox.ac.uk

variables simulated by individual members of that ensemble to be consistent with observations. A forecast that gets the warming rate wrong to 2020 is likely to continue getting it wrong to 2050. A forecast that underestimates the magnitude of an extreme El Niño event in temperature is likely also to underestimate it (after correcting for model bias) in precipitation. In both cases, there is still, in effect, only a single (albeit high-dimensional) point of verification.

This verification problem applies not only to forecasts of anthropogenic climate change but also to forecasting an extreme El Niño event or even a 100-year storm. Because different aspects of model physics, and hence different model errors, may be relevant in the simulation of events of different magnitudes, a probabilistic forecasting system that performs well at forecasting "normal" storms or "normal" El Niños may prove incapable of even simulating, let alone forecasting, a 100-year storm or 1000-year El Niño. But these are precisely the cases in which there is most pressure on the forecaster to "get it right" or at the very least to provide the user with some warning that conditions have reached the point where the forecast can no longer be relied upon.

This lack of any conceivable objective verification/falsification procedure has led some commentators to conclude that forecasts of anthropogenic climate change are fundamentally subjective and can never represent anything other than a formal expression of the beliefs of the forecaster(s). If you buy this view, then any probabilistic forecast of an unprecedented event goes down as subjective as well, so any El Niño forecaster with strongly-held beliefs who specialises in probabilistic forecasts of unusual El Niños (which, because they are unusual, cannot be subjected to traditional out-of-sample verification procedures) could claim equality of status with ECMWF. Worse still, assessment of forecasts of anthropogenic climate change degenerates all too easily into a dissection of the prior beliefs and motivations of the forecasters, "placing climate forecasts in their sociological context."

As die-hard old-fashioned realists, we firmly reject such New Age inclusivity. We also recognise, however, that the subjectivists have a point, in that conventional verification/falsification procedures cannot be applied to probabilistic climate forecasts, so it is up to us to clarify what we mean by a probabilistic forecast being "correct" or (a better-posed and generally more relevant question) more likely to be correct than the competition. The solution, we argue, is to focus on whether or not a forecast is likely to have *converged* in the sense that further developments in modelling (increasing model resolution, or including different parameterisations) are unlikely to result in a substantial revision of the estimated distribution of the forecast variable in question.

Modellers may find this counterintuitive, because they are accustomed to expecting an increase in model resolution or improved parameterisations to change the behaviour of their models in forecast mode, "otherwise why bother?" The point is that if the next generation of models change the forecast, what guarantee do we have that the generation after that will not change it again? Only a few years ago we were told that climate forecasts "could only be taken seriously" if performed with models that could run without surface flux adjustment. Now we are told they can only be taken seriously when they resolve oceanic eddies. But some aspects of simulated climate change have altered remarkably little, notably the overall shape (not the magnitude, which depends on the individual models' sensitivities) of the global mean response to observed and projected greenhouse gas emissions over the past and coming half-centuries. If all the changes in resolution, parameterisations or model structure that we attempt *fail* to alter some aspect of the forecast (the ratio of warming over the past 50 years to projected warming over the next 50 years under a specific emissions scenario, in this example), then we might hope that this aspect of the forecast is being determined by a combination of the observations and the basic physical constraints which all properly-formulated climate models share, such as energy conservation, and not by the arbitrary choices made by model developers.

We cannot, and will never be able to, treat model error in the formal manner in which observational error is taken in to account in short-range ensemble forecasting, because of the problems of sampling "all possible

models" and defining a metric for changes to model parameters and structure. The aim, therefore, must be to render this kind of model error irrelevant, by ensuring that, as far as possible, our forecasts depend on data and on the constraints that all physically conceivable models share rather than on any specific set of models, no matter how these are chosen or weighted. Such forecasts might be called STAID, or STAble Inferences from Data. STAID forecasts are unexcitable, largely immune from the whims of modelling opinion. They are also less sexy and attractive than forecasts based on a single (preferably young, but never free) super-high-resolution model. But ultimately, they are reliable: they will not change except through the painstaking acquisition and incorporation of new data.

In the final section of this article we lay out a methodology for STAID probabilistic forecasting by looking for convergent results across nested ensembles. Our focus will be on forecasting long-term climate change because this is the problem in which our current lack of a systematic treatment of model error is most acute, but the underlying principle could be applied to shorter-term (seasonal or even, ultimately, synoptic timescale) forecasting problems if the computing power becomes available for the necessary very large ensembles.

2 Model shortcomings, random errors and systematic errors

Any discussion of the model error in the context of weather and climate forecasting is immediately complicated by the fact that the term means different things to different people, so we will begin by setting out what we do not mean by model error. To the majority of model developers, "model error" evokes issues such as the fact that such-and-such a climate model does not contain a dynamical representation of sea-ice, or generates an unrealistic amount of low-level cloud. Their solution to these types of error is, naturally enough, to improve the model, either by increasing resolution or by introducing new physics. We will refer to this kind of research as resolving model shortcomings rather than a systematic treatment of model error, and it is not the focus of this article. While it is clearly desirable, is important to recognise that the process of resolving model shortcomings is open-ended and will never represent the whole story. No matter how high the resolution or detailed the physical parameterisations of models in the future, results will always be subject to the two further types of error: random errors due to the cumulative impact of unresolved processes on the resolved flow, and systematic errors due either to parameters not being adequately constrained by available observations or to the structure of the model being incapable of representing the phenomena of interest.

Although an important area of current research, the treatment of random errors is relatively straightforward at least in principle, and also not our focus here. Parameterisation schemes typically represent the impact of small-scale processes as a purely deterministic relationship between inputs and outputs defined in terms of the large-scale flow. Recently, experimentation has begun with explicitly stochastic parameterisation schemes as well as explicit representation of the effects of unresolved processes through stochastic perturbation of the physics tendency in the model prognostic equations [6, 30]. The theory of "stochastic optimals" [23, 26] is being developed as a means of identifying those components of stochastic forcing that contribute most to forecast error.

Whether or not stochastic forcing is included in the model, random unresolved processes will introduce unpredictable differences between model simulations and the real world, but this source of model error is relatively straightforward to treat in the context of linear estimation theory. A much more challenging problem is the treatment of systematic error, meaning those model biases that we either do not know about or have not yet had a chance to address.

Many model developers view proposals to develop a quantitative treatment of systematic error with sus-

picion since our objective appears to be to work out how to "live with" model shortcomings that remain unresolved, unexplained or simply unknown. Their concern is that if usable forecasts can be made with existing models, warts and all, then the case for further model development, increasing resolution and so forth may be weakened. As it turns out, a comprehensive treatment of systematic model error demands very substantial model development resources in itself, so this is not a realistic threat. There is an issue of the appropriate allocation of resources between quantifying the possible impact of systematic errors on the forecast system and attempting to get rid of them by improving the model. For some sources of error, it may be cheaper to eliminate them than to quantify their impact on the forecast, but this will never be true in all cases. Particularly on longer (climate) forecasting timescales, or in forecasting unprecedented events such as a record-breaking El Niño, we may simply not have the data or understanding to pin down crucial uncertainties in the model until after the event has occurred. For many of the most interesting forecasting situations, therefore, a systematic treatment of model error is essential for forecasts to be useful at all. Since no-one would suggest that we should keep our probabilistic forecasts useless so that we can maintain the case for making them less useless at some unspecified time in the future, some systematic treatment of model error is essential.

3 Analysis error & model error: helpful analogy or cul-de-sac?

The most obvious starting point for a treatment of model error in weather and climate forecasting is as an extension of the well-established literature on the treatment of analysis error in ensemble weather forecasting. This will have been discussed extensively elsewhere in these proceedings, so we only provide a cursory summary, and refer the reader to, for example, ref. [22], for more details and to ref. [25] for the extension of these principles to the multi-model context. Suppose the analysis from which a forecast is initialised is based on a standard optimal interpolation or Kalman filter, ignoring for present purposes the many technical issues regarding how such a filter might be implemented in practice:

$$\mathbf{x}_a = \mathbf{x}_b + (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}_b) \tag{1}$$

where $\mathbf{x}_a$ is the estimated state of the system at the time of the forecast initialisation, or more generally a four-dimensional description of the system over some period running up to the forecast initialisation time, $\mathbf{x}_b$ is the "background" or *a priori* estimate of the state vector $\mathbf{x}$ obtained, in a Kalman filter, by integrating the forward model from the previous analysis, $\mathbf{y}$ a set of observations that depend on the true state $\mathbf{x}$ via $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{u}$, $\mathbf{R}$ is the measurement noise covariance, $\mathbf{R} = <\mathbf{u}\mathbf{u}^T>$ (hence, for simplicity, incorporating error in the measurement operator $\mathbf{H}$ into $\mathbf{u}$) and $\mathbf{B}$ is the all-important, and typically rather obscure, "background error covariance" into which we might hope to bury our treatment of model error.

The derivation of the Kalman filter equations is based on the assumption that the evolution of the state vector $\mathbf{x}$ and the measurement error $\mathbf{u}$ can be treated, at least at the level of analysis uncertainty, as linear stochastic processes whose properties are completely determined by the covariance matrices $\mathbf{B}$ and $\mathbf{R}$, and that these matrices are known, or at least knowable: see ref. [17] for a more complete discussion. For a discretizable system in which

$$\mathbf{x}_t = \mathbf{M}\mathbf{x}_{t-1} + \mathbf{N}\mathbf{z}_t, \tag{2}$$

where $\mathbf{z}_t$ is a vector of unit-variance, uncorrelated noise, $<\mathbf{z}_t\mathbf{z}_t^T> = \mathbf{I}$ and the forward propagator $\mathbf{M}$ is known, then $\mathbf{B} = \mathbf{N}\mathbf{N}^T$. Note that $\mathbf{M}$ and $\mathbf{N}$ might be functions of $\mathbf{x}$, so the theory applies to non-linear models, provided $\mathbf{M}$ and $\mathbf{N}$ can be treated as constant over the analysis increment, $\mathbf{x}_a - \mathbf{x}_b$.

If $\mathbf{x}_b$ and $\mathbf{u}$ are multivariate normal, the analysis covariance, $\mathbf{A} = <(\mathbf{x}_a - \mathbf{x})(\mathbf{x}_a - \mathbf{x})^T>$, is given most compactly by

$$\mathbf{A}^{-1} = (\mathbf{H}\mathbf{R}\mathbf{H}^T)^{-1} + \mathbf{B}^{-1} \tag{3}$$

Since the notion that measurement errors can be treated, in some coordinate system, as multivariate normal is not too far-fetched, their contribution to the total analysis error is at least conceptually straightforward, although the actual specification of **R** for a wide range of multi-variate observations presents clear technical challenges. If the state vector really were generated by a linear stochastic process and the governing equations of that process are known (never the case in practice), then the contribution of the background error covariance, **B**, is also straightforward. The covariance matrix defined by equation (3) defines a multidimensional ellipsoid surrounding the best-guess estimate of the state vector $\mathbf{x}_a$ within which the true state of the system might be expected to lie. If **z** is simply a vector of unit-variance, uncorrelated, Gaussian random numbers, then possible state vectors, consistent with this analysis, are given by

$$\mathbf{x}' = \mathbf{x}_a + \mathbf{A}^{\frac{1}{2}}\mathbf{z}, \tag{4}$$

where $\mathbf{A}^{\frac{1}{2}}$ could consist, for example, of the eigenvectors of **A**, arranged columnwise, each multiplied by the square root of the corresponding eigenvalue: so each $\mathbf{x}'$ contains a random perturbation along each of the eigenvectors of **A** scaled by the standard deviation of the analysis error in that direction.

So far so good, but these may sound like sufficiently restrictive conditions such that this theory could never apply to, for example, a weather forecasting situation, since the state vector of the atmosphere is not governed by linear stochastic dynamics. In a low-order, noise-free chaotic dynamical system such as the Lorenz [21] model, which evolves on a fractal-dimensional attractor, then neither $\mathbf{x}_a$ nor any perturbed version thereof given by equation (4) will, with probability one, lie on the attractor [14]. This might appear not to matter if the system is such that trajectories converge rapidly back onto the attractor, although it may cause practical difficulties if the impact of the initial conditions of the forecast being "unbalanced" (off the system attractor) persist for a significant length of time (or worse, render the forecast numerically unstable). For this reason, considerable care is taken to "initialise" members of an ensemble weather forecast to reduce the impact of unphysically ageostrophic terms in their initial conditions.

A more fundamental difficulty arises from the fact that the distribution of trajectories consistent with a given set of observations and the dynamics of the underlying attractor may be highly non-uniform across the ellipsoid defined by equation (4), as illustrated in figure 2 of ref. [14]. One might hope that this kind of information could be reflected in the background error covariance matrix **B**, but because the space occupied by the attractor is fractal, no possible coordinate transformation could convert the pattern of accessible trajectories into a multinormal distribution.

These problems with the application of linear estimation theory to highly idealised, low-dimensional dynamical systems may, however, give a misleadingly negative impression of its applicability to much higher-order systems such as weather or climate models. The reason is that small-scale processes such as cloud formation may introduce sufficient high-order "noise" into a weather model such that the range of accessible trajectories is effectively space-filling (at least in the space of "balanced" perturbations) over regions consistent with the observations. Let us suppose for the sake of argument that the analysis is sufficiently accurate that analysis errors can be treated as linear and the notion of an ellipsoid of possible state vectors, consistent with the observations, given by equation (4) at least makes sense.

The fact that we are treating the dynamics as linear stochastic on small scales in the immediate vicinity of $\mathbf{x}_a$ does not, of course, restrict us to linear stochastic dynamics on larger scales, such as the propagation of **x** over the forecast lead time:

$$\mathbf{x}_f = \mathscr{M}(\mathbf{x}_a) \tag{5}$$

Indeed, exploiting the non-linearity of the system over the forecast time lies at the heart of so-called "optimal" forecast perturbation systems such as singular vectors [28] and breeding vectors [38]. Consider a two-dimensional example: in a coordinate system in which the analysis error is uniform in all directions,

errors in the vertical direction grow over the forecast lead time while errors in the horizontal direction decay. Hence, for any perturbation, only its projection onto the vertical matters for the forecast spread.

In this two-dimensional system, the variance of an arbitrarily-oriented vector in the vertical direction is half that of a vector of the same expected length oriented specifically in the vertical. Since this variance fraction declines with the dimension of manifold to which the state vector $\mathbf{x}$ is confined, an arbitrarily-oriented perturbation on a weather or climate model might project only a tiny fraction of its variance in any given direction (for a balanced, initialised perturbation the discrepancy might be smaller, but still large). If variance in all other directions were simply to disappear over the course of the forecast without affecting what happens to the projection of the perturbation onto the directions in which errors grow, then this initial orientation would not matter: we simply have to ensure that perturbations have sufficient power in all directions to populate the n dimensional ellipsoid defined by equation (4). But since, for a weather forecasting model, n may be extremely large, the result would be that we would need very high total amplitude perturbations in order to ensure that their projection onto a small number of error-growth patterns gives a representative forecast spread, and the larger the perturbations, the more difficult it becomes to ensure they are sufficiently balanced for a stable forecast. Hence perturbations are confined to the directions on which errors might be expected to grow, on the basis of the singular vector or breeding vector analysis.

In principle, each perturbation $\mathbf{x}'$ should contain a random component consistent with the analysis error in each of the n' (mutually orthogonal) directions in which errors are expected to grow, scaled such that $\mathbf{x}'^T\mathbf{A}^{-1}\mathbf{x}' = n'$. This is complicated by the fact that n' is flow-dependent and poorly determined in a complex non-linear system. In practice, perturbations are applied to rather more than directions than necessary with somewhat higher amplitude (there is some ambiguity in the "correct" amplitude in any case because of uncertainty in $\mathbf{A}$) and the dynamics of the ensemble sorts out n' (components of perturbations in directions that don't grow don't matter).

As stated in the introduction, the inclusion of random model error into this overall framework should be relatively straightforward. Uncertainty in the forward propagator of the analysis model, equation (2), arising from unknown small-scale unresolved processes, could be represented simply as an additional source of variance, augmenting $\mathbf{B}$ in the estimated analysis error. Likewise, in the generation of the ensemble forecast, individual forecast trajectories could be perturbed with the introduction of some form of "stochastic physics" term to represent the effect of this small-scale noise on the overall ensemble spread. Stochastic physics is already used in the ECMWF ensemble prediction system [6], and has been shown to improve ensemble representativeness at the medium range [30]. On longer timescales, a substantial body of literature is developing around the concept of "stochastic optimals," meaning (in a nutshell) forcing perturbations that have maximal impact on forecast spread, analogous to the optimal perturbations on initial conditions identified by singular vectors [23].

While important issues remain to be resolved in the appropriate specification of a stochastic physics term and the derivation of stochastic optimal perturbations, this is not the most challenging class of model error from a theoretical point of view. If unresolved small-scale processes exist that have an impact on the large-scale flow, then including them in the analysis or forecast model is a necessary model improvement, not a systematic treatment of model error as we interpret the term here.

4 Parametric and structural uncertainty and the problems of a metric for model error

Suppose the forecast model $\mathscr{M}$ in equation (5) contains a single underdetermined parameter, p, the uncertainty (prior distribution) of which is known, so $\mathscr{M} = \mathscr{M}(p)$. For the sake of simplicity, let us assume this uncertainty only affects the forecast model and not the model used to generate the analysis. Generating an ensemble forecast allowing for this uncertainty is straightforward. If p has no impact on the dynamics of error growth in the model, then we simply take initial conditions from the region defined by equation (4) and propagate these forward in time using a set of possible forecast models generated by making random perturbations to the parameter p consistent with its prior distribution. Because, *ex hypothesi*, p is independent of $\mathbf{x}_a$ there is no need to increase the size of the ensemble significantly: we simply perturb p at the same time as we perturb the initial conditions.

Generalising this to a vector of underdetermined parameters, $\mathbf{p}$, is also simple enough, provided the prior distribution of $\mathbf{p}$, $P(\mathbf{p})$, is known: we sample the distribution of possible models $\mathscr{M}(\mathbf{p})$ using random perturbations conditioned on $P(\mathbf{p})$. If perturbations to the initial conditions, $\mathbf{x}'$, have been made by perturbing a random combination of singular/breeding vectors simultaneously, then random perturbations to $\mathbf{p}$ can again be treated just like perturbations to the initial conditions.

If parametric perturbations interact with the growth of initial condition errors, then, again in principle, they can be treated with a simple extension of the singular or breeding vector technology. The goal is now to identify joint perturbations on parameters and initial conditions which maximise error growth, or for which

$$||\mathscr{M}([\mathbf{x}', \mathbf{p}'])||_f \gg ||[\mathbf{x}', \mathbf{p}']||_a. \tag{6}$$

The crucial element in all this is that we have a distance measure or metric for parameter perturbations analogous to the analysis error covariance for initial condition perturbations. If the distribution of $\mathbf{p}$ is multi-normal and there is no interaction with $\mathbf{x}'$, then a logical distance measure for contribution of $\mathbf{p}$ to the total error is provided by the inverse covariance matrix,

$$||\mathbf{p}'||_a = \mathbf{p}'^T \mathbf{C}_\mathbf{p}^{-1} \mathbf{p}'. \tag{7}$$

If the distribution of $\mathbf{p}$ is not normal but known, then some form of non-Euclidean distance measure could be defined to ensure an unbiased sampling of "possible" values of $\mathbf{p}$ where "possible" is defined, crucially, *without* reference to the observations used in the analysis.

As soon as we begin to consider structural uncertainty, or uncertainty in parameters for which no prior distribution is available, then all this tidy formalism breaks down. Unfortunately, the most important sources of model error in weather and climate forecasting are of precisely this pathological nature. The fundamental problem is the absence of a credible prior distribution of "possible models", defined in terms of model structure and inputs, from which a representative sample can be drawn. In order to perform such a representative sampling, we need to know how far apart two models are in terms of their structure, and how can we possibly compare the "size" of a perturbation involving reducing the spatial resolution by a factor of ten versus introducing semi-Lagrangian dynamics without reference to model output?

Ref. [12] makes a useful distinction between "measurable" inputs (like the acceleration due to gravity) and "tuning" inputs (like the choice of numerical scheme used to integrate the equations) to a computational model. The treatment of measurable inputs is straightforward: distributions can be provided based on what is consistent with direct observations of the process in question, and provided these observations are independent of those used subsequently to initialise the ensemble forecast, everything is unproblematic. Unfortunately, many of the parameters that have the most impact on the behaviour of climate models do not correspond to directly measurable quantities (although they may share names, like "viscosity"): defining

an objective prior distribution for such tuning inputs is effectively impossible, since we have no way of comparing the relative "likelihood" of different perturbations to model structure. The solution [18, 8, 12] must be to make use of the fact that models make predictions of the past as well as the future, and we discuss how this can be used to get around the problem of a lack of a defensible prior on the tuning parameters in the final sections of this paper. We believe the practical approach we suggest here should fit nicely into the formalism proposed by ref. [12], although there is much to be done on the details.

The difficulty of defining prior distributions without reference to the observations used to initialise the ensemble is particularly acute on climate timescales where the number of independent observations of relevant phenomena are extremely limited. The point is important because a number of modelling centres are beginning to adopt the following approach to ensemble climate forecasting which we might call the "likelihood-weighted perturbed-physics ensemble." First a collection of models is obtained either by gathering together whatever models are available (an "ensemble of opportunity") or by perturbing parameters in one particular model over ranges of uncertainty proposed by experts in the relevant parameterised processes. Second, members of this ensemble are weighted by some measure inversely proportional to their "likelihood" as indicated by their distance (dissimilarity) from observed climate. Third, a "probabilistic forecast" is generated from the weighted ensemble. Problems with this approach are discussed in the following section.

5 The problem with experts is that they know about things

The use of "expert prior" information in the treatment of model error in climate forecasting is sufficiently widespread that we feel we should devote a section to the problems intrinsic to this approach. Lest it be thought that we have any problem with experts, we will end up using expert opinion to design our perturbations in the concluding sections, but we will introduce additional steps in the analysis to minimise the impact of any "double-counting" this might introduce.

The problem with a direct implementation of a likelihood-weighted perturbed-physics ensemble is that some of the observations are almost certainly used twice, first in determining the perturbations made to the inputs and second in conditioning the ensemble. The result, of course, is that uncertainties are underestimated, perhaps by a significant margin.

Take a very simple example to begin with: suppose, no matter what parameters are perturbed in a climate model, the climate sensitivity (equilibrium response to doubling CO_2) varies in direct proportion to the net cloud forcing (CF) in the model-simulated present-day climate which, in turn, is constrained by the available observations to be zero, with standard deviation $\sigma_{CF} = 4\mathrm{Wm}^{-2}$ (if only things were so simple, but we simply wish to make a point here about methodology). We assemble an ensemble of models, either by collecting them up or by asking parameterisation developers to provide "credible" ranges for parameters in a single model. Suppose we find that the model-developers have done their homework well, and net cloud forcing in the resulting ensemble is also $0 \pm 4\mathrm{Wm}^{-2}$. We then weight the ensemble by $\exp(-CF^2/2\sigma_{CF}^2)$ to mimic their "likelihood" with respect to current observed climate and find that the variance of CF in the weighted ensemble is $\sigma_{CF}^2/2$.

If the model-developers had no knowledge of the fact that the perturbations they were proposing might have an impact on cloud forcing, or no knowledge of current observations and accepted uncertainty ranges for cloud forcing, this would be the correct result: if we double the number of independent normally-distributed pieces of information contributing to the same answer, then the variance in the answer is halved. But is it really credible that someone working on cloud parameterisations in a climate model could operate without knowledge of current observations? In reality, of course, the knowledge that a perturbation would

be likely to affect cloud forcing and the knowledge of the likely ranges of cloud forcing consistent with the observations would have conditioned the choice of "reasonable" perturbations to some unquantifiable extent, so uncertainties would have been underestimated by an unquantifiable margin.

In the case of "ensembles of opportunity" obtained by assembling models from different groups, the situation is likely to be worse than this, since no-one particularly wants their model to be the 2-σ outlier in something as basic as cloud forcing, which 5% of models should be if their distribution of behaviour is to be representative of current uncertainty. Hence uncertainty estimates from "raw" unweighted ensembles of opportunity are likely to be biased low (see, for example, ref. [29]), and estimates from likelihood-weighted ensembles of opportunity would be biased even lower. To make matters even worse, it has been proposed [11] that models should be explicitly penalised for being dissimilar to each other, which would further exacerbate the low bias in uncertainty estimates resulting from the natural social tendency of modelling groups each to aspire to produce the "best-guess" model.

The cloud forcing case is hypothetical, but there are practical examples of such problems, particularly with ensembles of opportunity. The curve and top axis in figure 1, following figure 1 of ref. [2], show an estimate of the distribution of warming attributable to greenhouse gases over the 20th century based on the analysis of ref. [37]. Consistent with current expert opinion, the "best guess" greenhouse induced warming, at 0.8K, is slightly higher than the total warming over the 20th century, with the net effect of other forcings estimated to be negative, but with a broad and more-or-less symmetric range of uncertainty due to internal variability and the possible confounding effects of other signals.

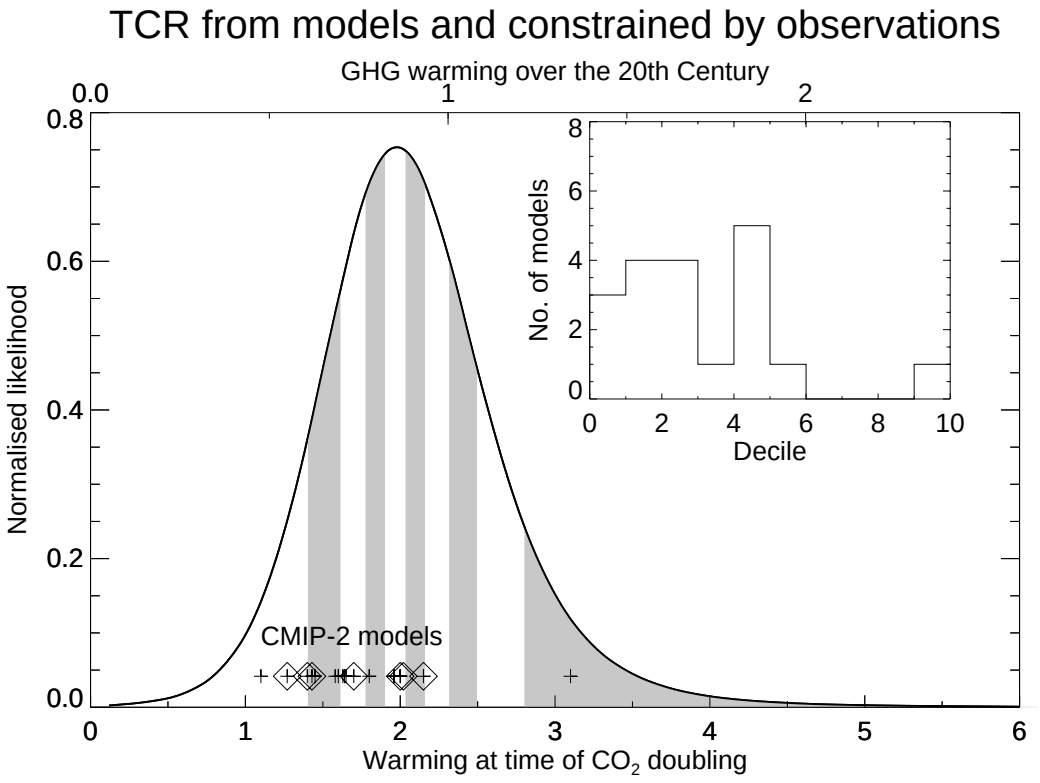

Figure 1: Comparison of the distribution of transient responses to increasing CO_2 expressed as attributable warming over the 20th century (top axis) and Transient Climate Response, TCR (bottom axis). Crosses indicate members of the CMIP-2 intercomparison, while diamonds show models included in the summary figures of the 2001 IPCC Scientific Assessment. Inset panel shows the number of CMIP-2 models falling into each decile of the distribution.

The bottom axis shows the same information expressed in terms of transient climate response (the expected warming at the time of CO_2 doubling under a 1% per year increasing CO_2 scenario). The advantage of TCR is that it has been calculated for a wide range of models in the CMIP-2 model intercomparison experiment [7], shown by the crosses on the plot. If these models were representative of behaviour consistent with observations, we would expect to see similar numbers falling in each decile (vertical stripe) of the distribution, and a more-or-less flat histogram in the inset panel. They clearly do not, with only two out of 19 models lying in the upper half of the distribution. Should we revise down our estimate of observed greenhouse-induced warming over the 20th century because the models generally under-predict it, or should we rather suspect that our sample of models is likely to be biased?

Problems also arise in the use of explicitly defined "expert priors", such as that used by ref. [10] (which we feel comfortable criticising since one of us was a co-author responsible for statistical methodology, although the author-of-intersection would like to emphasise, in self-defence, that this study was far from alone in its overall approach). In one of the cases considered, ref. [10] used an "expert prior" on climate sensitivity based on a systematic survey of the opinions of a number of climate experts to generate an initial sample of possible models. These were subsequently weighted by a goodness-of-fit statistic involving, among other things, the observed warming over the 20th century. This procedure is only valid if the opinions of the climate experts are independent of the observed warming.

There are climate experts who claim that they would still be of the opinion that the climate sensitivity (equilibrium warming on doubling of CO_2) is in the range 1.5-4.5K even if the world had not warmed up by 0.6K over the past century, more-or-less in line with a sensitivity somewhere near the middle of this range, and therefore this observation represents independent information. It is true that this range was originally proposed before much of the late-century warming took place, and also before the extent to which anthropogenic aerosol emissions might be masking CO_2-induced warming was established. Hence there might be something truth to the claim that it represented information independent of the observed warming trend back in 1979, but this is much harder to argue now: consider what the consensus view on climate sensitivity would be if the world had not warmed up at all, or had already warmed by 1.5K. Ideally, of course, we would like the opinion of experts in the physical processes represented in climate models who happen to be unaware of any of the observations we might subsequently use to condition the ensemble, but climate model development is an empirical science, so such "cloistered experts" (thanks to Lenny Smith for this nomenclature) are hard to find.

The problem becomes worse, not better, if the measure of model-data consistency is widened to encompass a broader range of observations. The more data are put into the ensemble-filtering step, the more likely it is that some of it was available to the experts designing the models or parameter perturbations used to generate the initial ensemble. If models are originally designed to fit the data, and then pruned to fit it better still, the resulting spread will seriously understate the likelihood that the particular cloud dataset that *everyone* uses to both tune and evaluate their cirrus parameterisation might just happen contain a fluke bias.

One response to this problem would be to exhort the modellers to try harder to produce models which are less like the observations in order to populate the tails of the distribution consistent with the datasets available. In the cloud forcing case, for example, if *CF* in the initial ensemble were uniformly distributed between $\pm 10 \mathrm{Wm}^{-2}$ then the application of likelihood-weighting gives almost exactly the correct variance for the weighted distribution. We have tried this approach in a seminar in one leading climate modelling centre, and would advise against it on safety grounds. On a purely practical level, models are typically designed such that they are numerically stable only in the vicinity of observed climate, so producing a truly "flat" prior distribution of something like *CF* would be impossible, and any convexity of the prior distribution across the range consistent with the observations would result in the variance of a likelihood-weighted ensemble being underestimated. The solution we propose in the final sections is simply to renormalise the

ensemble by the density of the prior distribution in the space of the observable in question, to mimic what we would have got had we been able to produce a flat prior, without actually going to the pain of producing one.

6 Empirical solutions and ensemble dressing: the probabilistic forecaster's flux-correction

If a succession of forecasts are available, based on the same forecasting system, of the same generic situation, then workable solutions to the problem of model error may be derived from a consideration of past verification statistics. Consider for simplicity a linear prediction system in which all errors are Gaussian. If it is found that the variance between the forecast verification and ensemble mean is systematically greater than the within-ensemble variance by a factor of two, the error can be corrected simply by inflating the forecast variance above that of the ensemble spread. Systematic forecast biases can also be corrected in this empirical way.

Much more sophisticated post-processing is required to account for non-linearity, discussed in Smith (these proceedings), but the principle is the same: the ensemble is treated as one source of information in a probabilistic forecast, along with many others. In some respects we might, at risk of annoying our co-presenters (always a desirable outcome), compare this kind of empirical "ensemble dressing" with the application of emprically-determined flux adjustments in climate models. It works, and turns an unusable ensemble forecasting system (or climate model) into something potentially highly informative, but at the same time it could seem disconcertingly ad hoc, and the question arises whether problems might be being papered over that would, if resolved, increase the utility of the forecast still further. Whatever our view on ensemble dressing, however, the fundamental problem is that it cannot be applied to what are often the most interesting forecasting situations: forecasting unprecedented or near-unprecedented events.

7 Probabilistic forecasts of one-off events: the verification problem

Empirical solutions may be highly effective in compensating for the impact of model error in generic forecasting situations for which a large number of verification instances can be collected. Problems arise when a forecast is to be made of an event which happens relatively seldom, or worse still is entirely unprecedented. The most obvious example of this nature is the problem of forecasting anthropogenic climate change, but similar problems arise on much shorter timescales as well. For example, in El Niño forecasting, only a couple of events have been observed in the kind of detail necessary to initialise a modern seasonal forecast, which is far too few to develop any kind of post-hoc corrections on ensemble spread. Moreover, each El Niño event evolves in a rather different way, making it likely that different sources of model error may be important in forecasting different events. Hence there is no guarantee that an empirical correction scheme, even if based on several El Niño events, will work for the next one.

The solution to this problem is intimately tied up in our basic understanding of what we mean by an "improved" or even "accurate" estimate of uncertainty in a forecast of the climate of 2050 or 2100. Assessing (and hence, in principle, improving) the accuracy of estimates of uncertainty in a probabilistic weather forecasting system is at least conceptually straightforward. We can examine a sequence of forecasts of a particular variable and keep track of how often the real world evolves into, say, the 5th percentile of the predicted distribution [9, 34]. If this occurs roughly 5% of the time, then the error estimates based on our forecasting system are acceptable. If it always occurs 10% of the time, then our error estimates can be re-calibrated accordingly. But how can we evaluate a probabilistic forecast of anthropogenic climate change

to 2050 when we will only have a single validation point, and that one available long after the forecast is of historical interest only?

This problem has led some commentators to assert [32] that any estimate of uncertainty in forecast climate change will always be fundamentally subjective, ultimately representing current scientific opinion rather than an objectively verifiable or falsifiable assertion of fact. While accepting that there will always be an element of subjectivity in the choice of methodological details or the choice of climate model(s) used in the analysis, we argue that this subjective element, including the dependence of results on the climate model(s) used, can and should be second-order, quantified and minimised as far as possible. To resign ourselves to any other position on an issue as contentious as climate change is to risk diverting attention from the science itself to the possible motivations of the experts or modelling communities on which current scientific opinion rests.

Although this sounds a relatively abstract issue, there are deeply practical consequences: the headline result of ref. [39] was that warming rates at the upper end of the range quoted in the IPCC 2001 Scientific Assessment [15] were extremely unlikely. This conclusion depended, to first order, on those authors' decision to assume (admirably clearly footnoted) that there was only a 1 in 20 likelihood of the climate sensitivity (equilibrium warming on doubling carbon dioxide) exceeding 4.5K, despite the fact that the available formal surveys of expert opinion [24] and estimates based on the analysis of climate observations [5, 10, 19, 13] suggest a substantially higher upper bound at this confidence level. Statements of the form "this is our opinion (or, perhaps less provocative but still problematic, our model), and these are its consequences" are unlikely to engender the kind of confidence in the non-scientific community required to justify far-reaching political and economic decisions.

How can we avoid the charge of subjectivism or "model-relativism"? The solution [3] is to focus on whether or not a probabilistic forecast for a particular climate variable has *converged*, rather than being side-tracked onto the unanswerable question of whether or not it is *correct*. In a similar vein (and, of course, not by coincidence), ref. [35] argues that we should look for consistency of results across classes of models.

A forecast distribution for a particular variable has converged if it depends primarily on the observations used to constrain the forecast and not, to first order, on the climate model used or subjective opinions of the forecasters. Hence, in claiming that a forecast distribution (of 2100 global temperature, for example, or Northwest European winter rainfall in the 2030s or, most challenging of all [31], some multivariate combination thereof) has converged, we are claiming that the forecast distribution is unlikely to change substantially due to the further development of climate models or evolution of expert opinion, although uncertainties are likely to continue to be reduced as the real-world signal evolves and more observations are brought to bear on the problem [37].

A claim of convergence is testable and falsifiable. For instance, it might be found that increasing model resolution systematically and substantially alters the forecast distribution of the variable in question, without the introduction of any new data. We can use our physical understanding of the system to assess whether this is likely to happen in any given instance, depending on the robustness of the constraints linking the forecast variable to observable quantities. In a complex non-linear system, however, any physically-based arguments will need to be tested through simulation.

8 Convergence of results across nested ensembles: a pragmatic approach to STAID forecasting

Once we have agreed on what what we are trying to do in the treatment of model error in a probabilistic forecast of a one-off event, there will doubtless be many different approaches to achieving this end. In these final two sections, we outline one approach and discuss the implications for the design of climate forecasting systems. In essence, our approach is inspired directly by the pragmatic approach taken for many years to model tuning. All models contain some tunable parameters which must be set initially at relatively arbitrary values and subsequently adjusted in the process of model development to optimise the fit between model output and observed climate. We simply propose extending this approach to optimise the fit between the distribution of models in a perturbed-physics ensemble and known uncertainties in observable properties of the system being modelled. The key difference is that when a model is being tuned the usual practice is to adjust the parameter in question and repeat the run, whereas when an ensemble is being tuned, it may be unrealistic to re-run it, so the weights assigned to ensemble members need to be adjusted instead.

What do we mean by "known uncertainties" and "observable properties"? We will focus on the "ideal" climate problem, in which the timescales of interest are much longer than the longest timescale of predictability in the system, so the role of the initial conditions can be ignored. A possible extension to problems in which initial conditions are important is discussed qualitatively in the final section. Suppose $\tilde{\mathbf{x}}_f$ is a particularly interesting quantity (externally driven global warming 2000-2030, for example) derived from the full description of the system at the forecast time, $\mathbf{x}_f$, and suppose also that we find, analysing a perturbed physics ensemble, that there is a consistent relationship between $\tilde{\mathbf{x}}_f$ and $\overline{\mathbf{x}}_a$, where $\overline{\mathbf{x}}_a$ is an observable property of the system over the analysis period (externally driven warming 1950-2000). We use the term "observable property" loosely to mean a quantity whose distribution can be constrained primarily by observations with only limited use of modelling, not necessarily something directly observable.

Let us represent all underdetermined "tunable inputs" [12] to the forecasting system for which we do not have a prior distribution, including structural uncertainties, as $\mathbf{q}$. Inputs for which a prior distribution is available that we can be sure is independent of the observations used in the forecast initialisation are treated like $\mathbf{p}$ above. $P(\mathbf{q})$ is undefinable because we have no way of sampling the space of all possible models and no way of saying how far apart two models are even if we could, but we need to assume some sort of distribution for $\mathbf{q}$, $\hat{P}(\mathbf{q})$, in order to get started. The crux of the solution is that we should design the forecasting system to minimise the impact of $\hat{P}(\mathbf{q})$ on the forecast quantity of interest, and only claim forecasts are STAID to the extent that they can be shown not to depend on $\hat{P}(\mathbf{q})$ (there is a direct analogy with the analysis of the role of the *a priori* in satellite retrievals).

The statement that there is a consistent relationship between $\tilde{\mathbf{x}}_f$ and $\overline{\mathbf{x}}_a$ which does not depend on the choice of model is tantamount to the claim that

$$P(\tilde{\mathbf{x}}_f|\overline{\mathbf{x}}_a,\mathbf{q}) = P(\tilde{\mathbf{x}}_f|\overline{\mathbf{x}}_a) \;\; \forall(\mathbf{q}). \tag{8}$$

This claim remains falsifiable even if it appears to be the case for all models (values of $\mathbf{q}$) tested to date, if the incorporation of a new process either changes the relationship between $\tilde{\mathbf{x}}_f$ and $\overline{\mathbf{x}}_a$ (sometimes called a transfer function [4]) or renders it dependent on new tunable inputs whose values are not constrained by observations. Such claims will be on strongest ground when they can be supported by a fundamental physical understanding of *why* there must be a consistent relationship between $\tilde{\mathbf{x}}_f$ and $\overline{\mathbf{x}}_a$ which any valid model must share, like energy conservation. Many, however, will remain "emergent constraints" on which models appear to have converged but are not necessarily dictated by the underlying physics.

Similarly, the claim that $\overline{\mathbf{x}}_a$ can be constrained primarily by observations is equivalent to the claim

$$P(\overline{\mathbf{x}}_a|\mathbf{y},\mathbf{q}) = P(\overline{\mathbf{x}}_a|\mathbf{y}) \;\; \forall(\mathbf{q}). \tag{9}$$

If there are no other constraints on the observing system, so $P(\mathbf{y})$ and $P(\overline{\mathbf{x}}_a)$ are both uniform, then $P(\overline{\mathbf{x}}_a|\mathbf{y}) = P(\mathbf{y}|\overline{\mathbf{x}}_a)$, meaning the likelihood of $\overline{\mathbf{x}}_a$ taking a certain value in the light of the observations can be equated with the likelihood of us obtaining these observations given that it does take that value. Whether or not it is reasonable to regard $P(\overline{\mathbf{x}}_a)$ as uniform over the range consistent with the observations will depend on the variable in question. For relatively well observed quantities like the underlying rate of global warming, this could be a reasonable assumption since, if models were to simulate warming rates consistently lower or higher than that observed, our response would be to revise the models or look for missing forcings, not revise our assessment of how fast the world was warming up (this is precisely what happened before the introduction of sulphate aerosol cooling in the early 1990s). The "attributable warming" shown in figure 1 is a slightly more derived quantity, but could still be regarded as a primary observable.

An estimate of the distribution of $\overline{\mathbf{x}}_a$ can be obtained from an ensemble of "hindcast" simulations:

$$\tilde{P}(\overline{\mathbf{x}}_a|\mathbf{y}) = \frac{P(\overline{\mathbf{x}}_a|\mathbf{y},\mathbf{q})\hat{P}(\mathbf{q}|\mathbf{y})}{\hat{P}(\mathbf{q}|\overline{\mathbf{x}}_a,\mathbf{y})} \simeq \frac{P(\mathbf{y}|\overline{\mathbf{x}}_a)\hat{P}(\mathbf{q}|\mathbf{y})}{\hat{P}(\mathbf{q}|\overline{\mathbf{x}}_a,\mathbf{y})} \tag{10}$$

where the second, more tentative, equality only holds when the above statements about $\overline{\mathbf{x}}_a$ being primarily constrained by observations are true. $\hat{P}(\mathbf{q}|\mathbf{y})$ represents a sample of models (values of the tuning inputs) obtained by perturbing parameterisations by collecting models and model-components from a range of different sources (an "ensemble-of-opportunity"). It is not an estimate of any distribution, because the distribution of all possible models is undefined, and (no matter what the experts claim) it is conditioned on the observations: we don't know how much, but our objective is to make it the case that we don't care. The denominator, $\hat{P}(\mathbf{q}|\overline{\mathbf{x}}_a,\mathbf{y})$ represents the frequency of occurrence of models in this sample in which the simulated $\overline{\mathbf{x}}_a$ lies within a unit distance from any given value, or the observed density of models in the space spanned by $\overline{\mathbf{x}}_a$, given the observations. Its role in equation (10) is rather like a histogram renormalisation in image processing: it ensures we end up with the same weight of ensemble members in each decile of the $P(\overline{\mathbf{x}}_a|\mathbf{y})$ distribution, forcing a flat inset histogram in figure 1.

Suppose $\mathbf{y}$ and $\overline{\mathbf{x}}_a$ each have only a single element, being the externally-forced global temperature trend over the past 50 years (figure 2). $\mathbf{y}$ is observed to be 0.15K per decade, $\pm$0.05K/decade due to observational uncertainty and internal variability. For simplicity of display, we assume internal variability is independent of $\mathbf{q}$, but relaxing this assumption is straightforward. The curve in figure 2 shows $P(\mathbf{y}|\overline{\mathbf{x}}_a)$ and the vertical lines show simulated $\overline{\mathbf{x}}_a$ from a hypothetical ensemble which, like the CMIP2 ensemble shown in figure 1, is biased with respect to the observations in both mean and spread. For simplicity, we have assumed a large initial-condition ensemble is available for each model-version (value of $\mathbf{q}$), so the $\overline{\mathbf{x}}_a$ from the models are delta-functions.

A likelihood-weighted perturbed-physics ensemble would estimate the distribution of $P(\overline{\mathbf{x}}_a|\mathbf{y})$ from

$$P'(\overline{\mathbf{x}}_a|\mathbf{y}) = P(\mathbf{y}|\overline{\mathbf{x}}_a)\hat{P}(\mathbf{q}|\mathbf{y}), \tag{11}$$

meaning weighting the ensemble members by some measure of their distance from observations and estimating the distribution from the result. This would only give the correct answer if $\hat{P}(\mathbf{q}|\overline{\mathbf{x}}_a,\mathbf{y})$ is uniform across the range consistent with the observations: note that this requires many more models that are almost certainly inconsistent with the observations than producing an ensemble that is consistent with the observations in the sense of providing a flat histogram in the inset panel of figure 1. In this case, of course, $\hat{P}(\mathbf{q}|\overline{\mathbf{x}}_a,\mathbf{y})$ is not uniform (the vertical lines are not uniformly distributed over the 0-0.3K/decade interval), so the likelihood-weighted perturbed-physics ensemble gives the wrong answer (dotted histogram).

A histogram-renormalised likelihood-weighted perturbed-physics ensemble (a bit of a mouthful, so let's just call it STAID) given by equation (10) gives the correct answer, but at a price. Because we need to

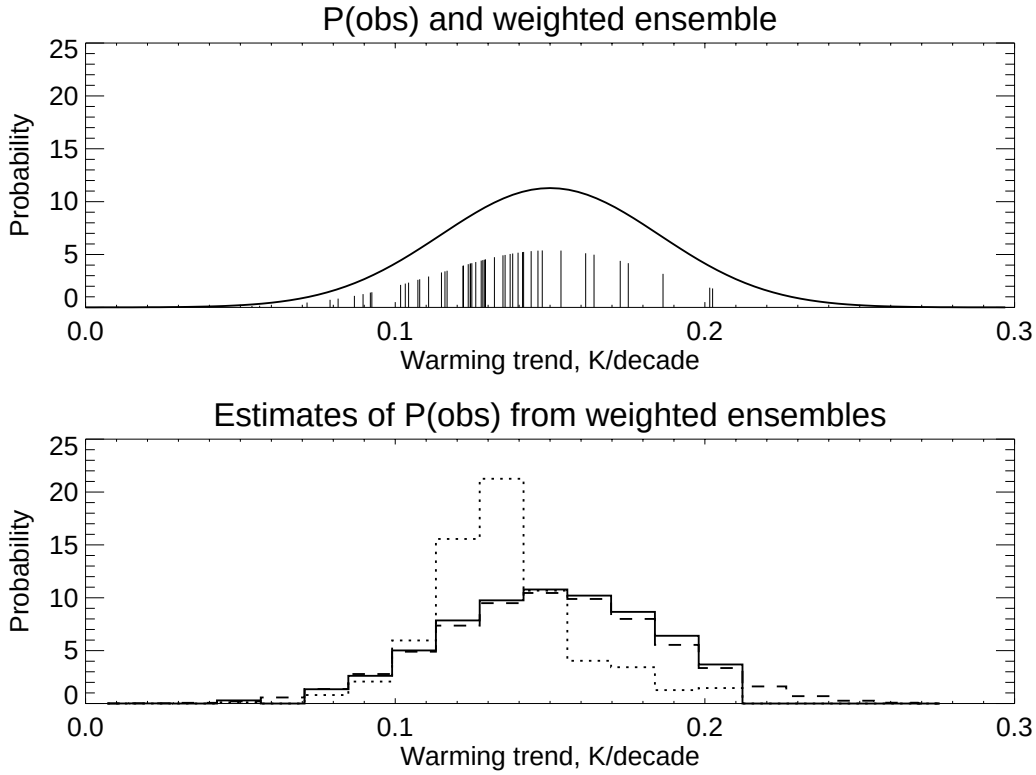

Figure 2: Top panel, curve: schematic distribution of externally-forced trends consistent with an observed trend of $0.15(\pm 0.05)$*K/decade,* $P(\overline{\mathbf{x}}_a|\mathbf{y})$*. Locations of vertical lines show model simulations from a hypothetical untargetted ensemble,* $\hat{P}(\mathbf{q},\mathbf{y})$*, with heights of lines showing likelihood-based weights* $P(\mathbf{y}|\overline{\mathbf{x}}_a)$*. Solid and dashed lines show estimates of* $P(\overline{\mathbf{x}}_a|\mathbf{y})$ *based on equation (10) with ensembles of 50 and 10,000 members respectively, while dotted histogram shows an estimate based on equation (11), which is biased in both mean and spread.*

compute the density of models in the space spanned by the observable quantity $\overline{\mathbf{x}}_a$, we need to ensure we have a large enough ensemble to populate this space. Sophisticated density-estimation methods would help (we have used a very unsophisticated one), but these would depend on assumptions about the smoothness of the response to variations in tunable inputs that could only be tested by simulation. The problem is that, for a given level of smoothness, the requisite ensemble size increases geometrically with the rank of $\overline{\mathbf{x}}_a$. In this case, $\overline{\mathbf{x}}_a$ has only a single element, and estimates $\hat{P}(\overline{\mathbf{x}}_a|\mathbf{y})$ from equation 10 are shown by the solid and dashed histograms in the figure, using 50 and 10,000-member ensembles respectively. While both are substantially closer to the true uncertainty range, there are significant distortions in the tails of the distribution (which are relevant to many policy decisions) estimated from the smaller ensemble. If no members of the ensemble happen to lie in a particular region consistent with recent climate observations, and we aren't justified in interpolating from adjacent regions, then no amount of re-weighting can help establish what would happen if they did. So STAID forecasting needs large ensembles. The key advantage of the STAID ensemble, however, is that its dependence on $\hat{P}(\mathbf{q})$ is second-order. Provided the ensemble is big enough to populate the space spanned by $\overline{\mathbf{x}}_a$, the impact of that troublesome prior is integrated out.

So far, this all looks like a very laborious way of recovering $P(\overline{\mathbf{x}}_a|\mathbf{y})$ which we have already said we are prepared to equate with $P(\mathbf{y}|\overline{\mathbf{x}}_a)$, so why bother? We are interested in $\overline{\mathbf{x}}_a$ because it is consistently related

to some forecast quantity, $\tilde{\mathbf{x}}_f$, independent of both $\mathbf{q}$ and $\mathbf{y}$. Hence

$$\tilde{P}(\tilde{\mathbf{x}}_f|\mathbf{y}) = \frac{P(\tilde{\mathbf{x}}_f|\overline{\mathbf{x}}_a)\tilde{P}(\overline{\mathbf{x}}_a|\mathbf{y})}{P(\overline{\mathbf{x}}_a|\tilde{\mathbf{x}}_f)} \quad \forall(\mathbf{q},\mathbf{y}), \tag{12}$$

where the transfer function $P(\tilde{\mathbf{x}}_f|\overline{\mathbf{x}}_a)/P(\overline{\mathbf{x}}_a|\tilde{\mathbf{x}}_f)$ is provided, ideally, by fundamental physics or, in practice, by the emergent constraints indicated by the ensemble. Having determined the weights on ensemble members necessary to recover an unbiased estimate of $P(\overline{\mathbf{x}}_a|\mathbf{y})$ and established that the relationship between $\tilde{\mathbf{x}}_f$ and $\overline{\mathbf{x}}_a$ is consistent across the ensemble and not dependent on $\mathbf{q}$, we simply apply those same weights to the forecast ensemble to arrive at an estimate of $P(\tilde{\mathbf{x}}_f)$ which is not, to first order, dependent on $\hat{P}(\mathbf{q})$. In a shorter-term forecasting context, ref. [16] note the importance of normalising out the prior when conditioning an ensemble with probabilistic information. Ref. [16] are conditioning a climatological timeseries model of local weather using a probabilistic forecast whereas we are conditioning a perturbed-physics ensemble using probabilistic information about ranges of past warming rates consistent with recent observations, but the underlying objective is the same: a smooth progression of estimated distributions, making maximal use of available information, throughout the forecast period.

Renormalising by prior density is a simple enough manouver, required by a 300-year-old theorem, but it has very profound implications for experimental design. Very few forecast quantities of interest will be found to depend exclusively on only a single observable climate variable. That said, given the strong constraints linking different variables in climate models, the number of effectively independent observable dimensions might be relatively small (fewer than half-a-dozen) at least for large-scale forecast quantities. Sampling a four-dimensional space at decile resolution for the computation of $P(\mathbf{q}|\overline{\mathbf{x}}_a,\mathbf{y})$ requires, however, an $0(10^4)$-member ensemble of climate models, and if we also allow for initial-condition ensembles to increase the signal-to-noise and boundary-condition ensembles to allow for uncertainty in past and future forcing, the desired ensemble size runs into millions. Fortunately, such ensembles are now feasible using public-resource distributed computing [1, 36, 3].

Interestingly, if we give up the chase for a defensible prior encompassing model error, we no longer have any universal "best" set of weights to apply to a perturbed-physics ensemble: the weights will depend on the forecast quantity of interest. The reason is that they are determined by whatever it takes to make the ensemble consistent with current knowledge and uncertainty in observable quantities on which that forecast quantity is found to depend, and different forecast quantities will depend on different observables. There is nothing inconsistent about this, since model errors are likely to have a different impact on different variables, but it does mean that forecast applications need to be integrated much more closely into the forecasting system itself. If a climate impact, for example, depends on some combination of temperature and precipitation, we cannot simply combine results from a weighted ensemble targetting temperature with another targetting precipitation: we need to recompute the weights to identify the observable variables that specifically constrain the function of interest. The good news, however, is that provided the initial ensemble is big enough to be space-filling in the relevant observables, then this is simply a post-processing exercise which does not require us to re-run the models.

9 Practical implications for weather and climate forecasting

Consider the following scenario: a small number of members of a perturbed-physics or multi-model ensemble indicate a storm of unprecedented magnitude striking Paris in a couple of days time. Examination of the ensemble statistics indicate that the magnitude of this storm is strongly correlated with the depth of a depression now forming off Cape Hatteras, independent of the model considered or perturbation made to model physics. Of course, the immediate response is to scramble our adaptive observation systems to get

more data on this depression, but all available adaptive observation systems have been re-deployed to some trouble-spot in a remote corner of the globe. Worse still, there is a technical problem with the computer, the system manager is on a tea-break (inconceivable at ECMWF, but for the sake of argument), so there is no chance of re-running any members of the ensemble. The mayor of Paris has heard rumours in the press and is on the line demanding an estimate of the chance his roof is going to get blown off. Sounds tough? Welcome to climate research.

Suppose most members of the ensemble display a depression off Cape Hatteras which is consistent with the depth of the observed depression but 90% of the model depressions, $\overline{\mathbf{x}}_a$, are weaker than the observed best-guess depth $\mathbf{y}$. If we adopted a likelihood-weighted perturbed-physics approach, we would simply weight by the distance from observations (independent of sign), which would have the desirable effect of downweighting ensemble members whose depressions are much weaker than observed, but would leave a bias between the ensemble-based estimate of the depth of the depression and a model-free observation-based estimate. If we were very confident in our prior selection procedure for the inclusion of models into our ensemble, then this bias is desirable: the ensemble system is telling us the observations are, more likely than not, overestimating the depth of the depression. If, however, as this article has argued, we can never have any confidence in a prior selection procedure that purports to encompass model error, then we certainly shouldn't be revising the observations in the light of any model-based prior. Instead, equation (10) would imply we should renormalise the ensemble histogram to give equal weight to the 90% of members that underpredict the depression (and hence, if the relationship is monotonic, are likely to underpredict the storm) as to the 10% that overpredict: bad news for the Mayor.

Crucially, a rival forecasting centre which finds the same relationship between depression and storm but whose prior model selection is such that 90% of ensemble members overpredict the depth of the depression would, if also using equation (10) to apply a histogram renormalisation, give the Mayor the same message: good news for the scientific reputation of the forecasters. The storm is unprecedented, so neither forecast is verifiable in the conventional sense, but we can at least make sure they are STAID.

This kind of reassessment of probabilities in the light of expert judgment about the origins of biases goes on all the time already, and in time it is conceivable to imagine extended-range forecasting systems in which multi-model perturbed-physics ensembles are searched automatically for observable precursors to forecast events of interest, and then reweighted to ensure the space spanned by these observables is representatively populated. Tim asked us to discuss how the treatment of model error might apply to shorter-timescale problems than the climate timescale we normally deal with, and we've had a go, conscious there is a substantial literature on these shorter-term issues with which we are only vaguely familiar (apologies in advance to all the relevant authors we have failed to cite). In the shorter term, however, the ideas proposed in this article are likely to have much more relevance to the climate research problem, where we have the time and resources to do the problem properly, and hence no excuse not to.

The conventional climate modelling approach (tuning a maximum-feasible-resolution model to obtain an acceptably stable base climate, followed by a sequence of simulations using only a single version thereof) provides no new information unless the new model's forecasts happen to lie completely outside the forecast distribution consistent with current observations based on a lower resolution set of models. Given the large number of underdetermined parameters in even the highest-resolution models, the question of whether a single version of a high-resolution model displays a larger or smaller response to external forcing than previous models is vacuous. The interesting question, in the context of probabilistic climate forecasting, is whether the increase in resolution has systematically altered the range of behaviour accessible to the model under a comprehensive perturbation analysis, which cannot be addressed for a model being run at the maximum resolution feasible for a single integration or initial-condition ensemble.

The challenge is to use climate models to identify constraints on the response to various possible forcing

scenarios: i.e. using models to establish what the climate system cannot (or is unlikely to) do, a much more demanding task than using them to describe some things it might do. Such constraints are likely to be fuzzy so, for many variables of interest, the spread of responses in currently-available models may be too small to determine them effectively, as was the case for the CMIP-2 results, taken alone, in figure 1. Large ensembles of models with deliberately perturbed physical parametrisations may be needed to provide enough diversity of response to identify these constraints. Adjoint-based optimal perturbation techniques used in shorter-range ensemble forecasting [27] may not be extendable, even in principle, to the climate problem [20], further increasing the size of ensembles required.

If constraints can be identified that, on physical grounds, we might expect any correctly-formulated climate model to share, then inferences based on these constraints are likely to be robust. Some constraints will be obvious (like energy conservation) but these will typically not be enough to constrain any useful forecast quantities. Others will be more subtle and open to falsification, like the constraint that the sensitivity parameter (a measure of the net strength of atmospheric feedbacks) does not change in response to forcings up to 2-3 times pre-industrial CO_2. This is a property shared by almost all climate models available to date but it could, in principle, be falsified if higher-resolution models were to display *systematically* non-linear sensitivities. Hence a probabilistic forecast of a particular variable that depends on the linearity of atmospheric feedbacks is more open to falsification than one that depended solely on energy conservation, which brings us round full-circle. We will depend on the expert judgement of modellers to assess the reliability of our probabilistic forecasts (the likelihood that they will be falsified by the next generation of models), but, in a crucial step forward, our reliance on expert judgement will be second-order. We need experts to assess the reliability of our uncertainty estimates rather then to provide direct input into the uncertainty estimates themselves.

Acknowledgements This article represents a synthesis of (we hope ongoing) conversations over a number of years with, among many others, Mat Collins, Dave Frame, Jonathan Gregory, William Ingram, John Mitchell, James Murphy, Catherine Senior, David Sexton, Peter Stott, Simon Tett, Alan Thorpe and Francis Zwiers. Afficionados will recognise many of Lenny Smith's ideas in here, more than we can acknowledge specifically. We are grateful to Michael Goldstein and Jonathan Rougier for making ref. [12] available, to Tim Palmer for his invitation and good-humoured response to the talk, and to Els Kooij-Connally for her forebearance in the preparation of these proceedings.

References

[1] M. R. Allen. Do-it-yourself climate prediction. *Nature*, 401:642, 1999.

[2] M. R. Allen and W. J. Ingram. Constraints on future climate change and the hydrological cycle. *Nature*, 419:224–232, 2002.

[3] M. R. Allen and D. A. Stainforth. Towards objective probabilistic climate forecasting. *Nature*, 419:228, 2002.

[4] M. R. Allen, P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. Delworth. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, 407:617–620, 2000.

[5] N.G. Andronova and M.E. Schlesinger. Causes of global temperature changes during the 19th and 20th centuries. *Geophysical Research Letters*, 27:2137–3140, 2000.

[6] R. Buizza, M. Miller, and T. N. Palmer. Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 125B:2887–2908, 1999.

[7] C. Covey, K. M. AchutaRao, S. J. Lambert, and K. E. Taylor. Intercomparison of present and future climates simulated by coupled ocean-atmosphere GCMs. Technical Report 66, PCMDI, 2000.

[8] P. S. Craig, M. Goldstein, J. C. Rougier, and A. H. Seheult. Bayesian forecasting for complex systems using computer simulators. *J. Am. Statist. Ass.*, 96:717–729, 2001.

[9] ECMWF. http://www.ecmwf.int/products/forecasts/guide/Talagrand diagram.html. 2002.

[10] C. E. Forest, P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, 295:113–117, 2002.

[11] F. Giorgi and L. O. Mearns. Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method. *Journal of Climate*, 15:1141–1158, 2002.

[12] M. Goldstein and J. Rougier. Probabilistic models for transferring inferences from mathematical models to physical systems. 2003. submitted & available from `http://maths.dur.ac.uk/stats/physpred/papers/directSim.pdf`.

[13] J. M. Gregory, R. Stouffer, S. Raper, N. Rayner, and P. A. Stott. An observationally-based estimate of the climate sensitivity. *Journal of Climate*, 15, 2002. 3117-3121.

[14] J. A. Hansen and L. A. Smith. Probabilistic noise reduction. *Tellus*, 5:585–598, 2001.

[15] J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson (eds.). *Climate Change 2001: The Science of Climate Change*. Cambridge University Press, 2001.

[16] S. P. Jewson and R. Caballero. The use of weather forecasts in the pricing of weather derivatives. *Meteorological Applications*, 2003. Submitted & available on `www.ssrn.com`.

[17] K. Judd. Nonlinear state estimation, indistinguishable states and the extended kalman filter. *Physica D*, To appear, 2003.

[18] M. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. Roy. Statist. Soc. B*, 63:425–464, 2001.

[19] R. Knutti, T.F. Stocker, F. Joos, and G.K. Plattner GK. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature*, 416:719–723, 2002.

[20] D. J. Lea, M. R. Allen, and T. W. N. Haine. Sensitivity analysis of the climate of a chaotic system. *Tellus*, 52A:523–532, 2000.

[21] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmos. Sci.*, 20:130–141, 1963.

[22] F. Molteni, R. Buizza, T. N. Palmer, and T. Petroliagis. The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122A:73–119, 1996.

[23] A. M. Moore and R. Kleeman. Stochastic forcing of ENSO by the intraseasonal oscillation. *Journal of Climate*, 12:1199–1220, 1999.

[24] M. G. Morgan and D. W. Keith. Subjective judgements by climate experts. *Environmental Policy Analysis*, 29:468–476, 1995.

[25] K. R. Mylne, R. E. Evans, and R. T. Clark. Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quarterly Journal of the Royal Meteorological Society*, 128:361–384, 2002.

[26] D. S. Nolan and B. F. Farrell. The intensification of two-dimensional swirling flows by stochastic asymmetric forcing. *Journal of Atmos. Sci.*, 56:3937–3962, 1999.

[27] T. N. Palmer. Predicting uncertainty in forecasts of weather and climate. *Rep. Progress in Physics*, 63:71–116, 2000.

[28] T. N. Palmer, R. Buizza, F. Molteni, Y. Q. Chen, and S. Corti. Singular vectors and the predictability of weather and climate. *Philosophical Transactions of the Royal Society of London*, A:348:459–475, 1994.

[29] T.N. Palmer and J.Raisanen. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature*, 415:512–514, 2002.

[30] K. Puri, J. Barkmeijer, and T. N. Palmer. Ensemble prediction of tropical cyclones using targeted diabatic singular vectors. *Quarterly Journal of the Royal Meteorological Society*, 127B:709–731, 2001.

[31] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660, 2002.

[32] S. H. Schneider. Can we estimate the likelihood of climatic changes at 2100? *Climatic Change*, 52, 2002.

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

[34] L. A. Smith. Disentangling uncertainty and error: on the predictability of nonlinear systems. In Alistair I. Mees, editor, *Nonlinear Dynamics and Statistics*, pages 31–64. Birkhauser, 2000.

[35] L. A. Smith. What might we learn from climate forecasts? *Proceedings of the National Academy of Sciences*, 99:2487–2492, 2002.

[36] D. Stainforth, J. Kettleborough, M. Allen, M. Collins, A. Heaps, and J. Murphy. Distributed computing for public-interest climate modeling research. *Computing in Science and Engineering*, 4:82–89, 2002.

[37] P. A. Stott and J. A. Kettleborough. Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, 416:723–726, 2002.

[38] Z. Toth and E. Kalnay. Ensemble forecasting at ncep and the breeding method. *Monthly Weather Review*, 125:3297–3319, 1997.

[39] T. M. L. Wigley and S. C. B. Raper. Interpretation of high projections for global-mean warming. *Science*, 293:451–454, 2001.