Project no. GOCE-CT-2003-505539

Project acronym: ENSEMBLES

Project title: ENSEMBLE-based Predictions of Climate Changes and their Impacts

Instrument: Integrated Project

Thematic Priority: Global Change and Ecosystems

D1.12 A report/publication comparing the Oxford and Hadley Centre methods for obtaining probabilistic climate forecasts from perturbed parameter ensembles.

Due date of deliverable: February 2008
Actual submission date: December 2009

Start date of project: 1 September 2004 Duration: 60 Months

Organisation name of lead contractor for this deliverable: UOXFDC

Revision [draft, 1, 2, ..]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

Methods of weighting ensembles

Myles Allen, Department of Physics, University of Oxford

December 31, 2009

1 Summary and overview

Various schools of thought have emerged over the course of the ENSEMBLES project regarding the weighting and interpretation of ensembles, reflecting long-standing differences in the statistics community on the presentation of uncertainty. Two broad distinctions need to be made to relate various recent studies: differences between methods used to assign a likelihood or goodness-of-fit statistic to individual members of an ensemble, and differences in sampling methods used to generate the ensemble itself. In this report, we summarise the approaches used, illustrating that differences between them reflect fundamentally different objectives. Hence results from the different methods are in fact attempting to do different things, and so should not be expected to correspond to each other, although we should expect predictable inequality relationships between them.

Some of the earliest studies attempting to quantify uncertainty in climate forecasts emerged directly from the detection and attribution literature of the 1990s, notably the optimal fingerprinting approach of [*Hasselmann*, 1993, *Hasselmann*, 1997], [*Santer* et al., 1994] and [*Hegerl* et al., 1996]. [*Leroy*, 1998] and [*Allen & Tett*, 1999] observed that optimal fingerprinting could be cast as a linear regression problem in which it is assumed that climate models simulate the patterns of the climate response to various external drivers correctly, and observations are used to estimate the magnitude of that response. A subsequent generalisation by [**?**] allows for some uncertainty in the patterns of response, but is still based on the principle that models provide much more reliable information regarding response patterns than response magnitudes.

The physical justification for this principle is strong: the spatial pattern of response to, for example, greenhouse forcing is driven by the differences in heat capacity between land and ocean and the location of the continents, which are not model-dependent. Likewise, the temporal pattern of response depends primarily on the time-history of greenhouse forcing and only secondarily on the time-scales of the response. In contrast, the magnitude of the response depends on the transient climate response, or TCR. This in turn depends on the atmospheric feedbacks that control the equilibrium climate sensitivity and on the efficiency of ocean heat uptake, both of which are uncertain.

Hence, in the context of multi-model and "perturbed physics" ensembles, optimal fingerprinting is equivalent to generating a large "pseudo-ensemble" simply by taking the mean pattern of response to a given external forcing as simulated by a small ensemble and scaling it up and down by an arbitrary parameter representing uncertainty in response magnitude. It is important that responses to short-term (e.g. volcanic) and long-term (e.g. most anthropogenic) forcings are estimated separately using a multiple regression, since uncertainty in the time-constants of the climate system (primarily linked to ocean heat uptake) mean that errors in response magnitude may be very different on different timescales. Ideally, the response to anthropogenic aerosol forcing should also be estimated separately from the response to greenhouse forcing: although both operate on similar timescales, some potential sources of uncertainty in the aerosol response do not affect the greenhouse response, and vice versa. Hence a pre-requisite for this approach are separate simulations of the responses to individual forcings, either separately or in combinations.

The goodness-of-fit between individual members of this pseudo-ensemble are then evaluated with a standard weighted sum of squares, with the expected model-data differences due to internal climate variability, observation error and (in some studies) model pattern uncertainty providing the matrix of weights or metric. The range of, for example, the warming attributable to anthropogenic greenhouse gas increases over the past 50 years across the members of this pseudo-ensemble that fit the data better than would be expected by chance in, say, 90% of cases provides a confidence interval on this quantity. This approach is the primary information source for attribution statements in the IPCC Third and Fourth Assessments.

Applying the same scaling factors to model-simulated responses to future forcing provides a natural method of deriving confidence intervals on future climate change. This approach was used by [*Allen* et al., 2000], [*Stott & Kettleborough*, 2002] and, for regional changes, by [**?**], and has been referred to as the ASK approach. The crucial assumption (which is also implicit in attribution studies) is that fractional errors in model-simulated responses persist over time, so a model that underestimates the past response to a given forcing by, for example, 30% may be expected to continue to do so in the future. This assumption is supported by comparing model results for scenarios under which forcing is sustained into the future, such as A1B [**?**], but [*Allen* et al., 2000] note that it would be less reliable for stabilisation scenarios.

The ASK approach can provide ranges of uncertainty in forecast climate that may, for variables that are poorly constrained by observations, be much wider than the range of available model simulations. This was clearly an advantage when very few models were available, and will continue to be necessary as long as the spread of model simulations is thought to underestimate the full range of uncertainty. ASK therefore provides a complementary approach to more recent methods of probabilistic forecasting such as weighted or un-weighted perturbed-physics or multi-model ensembles. There are, however, some important points of principle in which ASK as traditionally implemented differs from most ensemble-based approaches, which need to be addressed if results are to be compared

cleanly.

In contrast to the ASK approach, [*Murphy* et al., 2004, **?**, **?**] adopt an explicitly Bayesian approach, building on earlier work by [**?**]. Ensembles are generated by varying parameters using subjective assessments of parameter uncertainty and weighted by their goodness-of-fit to observations. Distributions that emerge from this approach have an explicit probabilistic interpretation as the degree of belief in the relative probability of different outcomes in the light of the evidence available. Consistent with the attribution literature, ASK provides classical ("frequentist") confidence intervals - that is, ranges over which models match observations better than a given threshold for goodness-of-fit. In contrast, most ensemble-based approaches provide Bayesian posterior probability intervals - ranges within which a given percentage of the weighted ensemble is found to lie. These are only comparable if ensemble members are distributed uniformly across the observable quantities that are used to constrain them and uncertainties in these quantities are approximately Gaussian (the so-called Jeffreys Prior condition). If the constraints provided by the observations are weak and models tend to cluster near the best-fitting model (as would be expected if all modelling groups are aiming to simulate observations as well as possible), these conditions are not satisfied, so ranges provided by ASK are not directly comparable to ranges provided by other approaches. Worse, ranges on forecast anthropogenic warming will then not be consistent with ranges on past anthropogenic warming, leading to the absurd conclusion that we are less uncertain about the future than we are about the recent past [**?**].

While the climate research community clearly cannot expect to resolve an issue that has dogged the entire statistics literature for decades, we can at least be clear which approach, classical or Bayesian, is being used in the presentation of uncertainty. A fundamental issue that needs to be addressed is that the standard uncertainty qualifiers used by Working Group 1 "likely", "very likely" etc.) are used to refer both to classical confidence intervals and Bayesian posterior probability intervals. The nominal definition of these qualifiers is unambiguously Bayesian, but in many, perhaps most, instances they are used to refer to classical confidence intervals or the results of hypothesis tests. This ambiguity of usage within IPCC has already attracted criticism among statisticians (Spiegelhalter, *pers. comm.* 2008). A simple solution would be to restrict the use of "likely" etc. to cases in which a confidence interval can be derived or hypothesis test performed (which refer, appropriately, to likelihoods of goodness-of-fit) and to use the more explicitly Bayesian language recommended by [**?**] for Bayesian posterior probabilities.

1.1 Analysis framework

All ensemble climate forecasting systems comprise, at some level, a modelling framework $\mathcal{M}$ which, given a set of input parameters Θ, generates a simulation $\mathbf{x}_\mathrm{o}$ of quantities that can be observed and a prediction $\mathbf{x}_\mathrm{f}$ of quantities that we wish to forecast:

$$(\mathbf{x}_\mathrm{o}, \mathbf{x}_\mathrm{f}) = \mathcal{M}(\Theta) \ . \tag{1}$$

If it is known that some input parameters (initial conditions in model-year 1860, for example) have a completely unpredictable impact on both observations and forecast, $\mathbf{x}_\mathrm{o}$ and $\mathbf{x}_\mathrm{f}$ may comprise the mean of an ensemble generated by a set of random settings of those parameters. For the sake of generality, we assume that $\mathbf{x}_\mathrm{o}$ is directly comparable to actual observations $\mathbf{y}$, meaning that the properties of the "measurement operator" linking model-simulated to observed quantities, representing for example the impact of incomplete sampling, are incorporated into $\mathcal{M}$. Furthermore, simulations must be set up in such a way such that the *expected irreducible* difference between models and observations is zero:

$$\min_{\Theta} \langle \mathbf{y} - \mathbf{x}_\mathrm{o} \rangle = \mathbf{0} \quad , \tag{2}$$

meaning that the minimum difference between models and observations obtained by varying Θ is centred on zero. This does not mean that the individual model simulations are unbiased, or even that the best available model is unbiased, but that the overall modelling framework $\mathcal{M}$ is unbiased in the sense that all known forcings and physical processes that are likely to have an impact on $\mathbf{x}_\mathrm{o}$ have been taken into account and known sources of bias have been removed prior to comparison with observations. If there is a known bias in model climatology, for example, this can be taken into account in the modelling framework $\mathcal{M}$ by expressing $\mathbf{y}$ and $\mathbf{x}_\mathrm{o}$ as anomalies about their respective time mean values.

An essential ingredient in an ensemble climate forecasting system is a prediction of how we expect this minimum model-data difference to be distributed about zero. Where these differences are dominated by internal climate variability and observational error, they are generally modelled as Gaussian, expressed in terms of a covariance matrix:

$$\mathbf{C} = \min_{\Theta} \langle (\mathbf{y} - \mathbf{x}_\mathrm{o})(\mathbf{y} - \mathbf{x}_\mathrm{o})^T \rangle \quad . \tag{3}$$

This matrix is also a model-predicted quantity, estimated for example from extended control simulations to estimate the properties of internal climate variability. The matrix $\mathbf{C}$ may also be augmented by a "discrepancy term" to represent irreducible model error, assuming this can also be represented as a zero-mean Gaussian quantity.

2 Metrics of individual model quality

All but the simplest "ensemble-of-opportunity" approaches to generating a range of uncertainty on a climate forecast require some measure of the quality of individual climate models or model-versions. In general, this can be characterised as a distance measure, often expressed as a weighted sum squared difference between a model simulation $\mathbf{x}$, which may be the mean of an initial-condition ensemble, and the corresponding set of observations $\mathbf{y}$:

$$r^2 = (\mathbf{y} - \mathbf{x}_\mathrm{o})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{x}_\mathrm{o}) \quad , \tag{4}$$

where $\mathbf{C}$ is a measure of the expected difference between model and observations due to processes that can be treated as random. Assuming the simulation generating $\mathbf{x}_o$ is not initialised from the observations, this is typically dominated by internal climate variability, but depending on the complexity of the analysis may also include a representation of observational error, forcing error, irreducible model error and so on.

Under the assumption that errors are Gaussian and that the distributions of $\mathbf{x}$ and $\mathbf{C}$ are determined by a set of parameters Θ, the discrepancy can be expressed as a likelihood:

$$\mathcal{L}(\Theta|\mathbf{y}) = \frac{1}{\sqrt{2\pi^n}} \exp\left(-\frac{r^2}{2}\right) \tag{5}$$

where n is the rank of $\mathbf{C}$, or the number of independent observations. In the case of ASK-type regression approaches, Θ is simply the parameters of the regression model, or undetermined scaling factors to be applied to model-simulated responses to individual forcing agents, while in the simplest perturbed-parameter ensembles, Θ represents the parameters perturbed in the climate model itself. The interpretation of Θ is more complicated when structural model uncertainty is taken into account, but for the sake of unity, we will assume that structural errors can be parameterised, noting that this assumption is controversial.

Absolute versus relative likelihoods

In a Bayesian analysis, the likelihood $\mathcal{L}(\Theta|\mathbf{y})$ is simply the probability density function of obtaining a simulation $\mathbf{x}_o$ in the vicinity of $\mathbf{y}$ given the parameters Θ:

$$\mathcal{L}(\Theta|\mathbf{y}) = \Pr(\mathbf{x}_o = y|\Theta) \ . \tag{6}$$

Clearly, this tends to become progressively smaller the higher the rank of $\mathbf{y}$ simply because the probability of the simulation "hitting the target" falls off the higher the dimension of the observation space considered. Hence the absolute likelihood of any setting of the parameters Θ depends, even for a structurally perfect model, on the number of observations used to constrain it, making the interpretation of absolute likelihoods rather obscure. Hence all studies rely more-or-less explicitly on relative likelihoods. The relative likelihood of two sets of parameters, Θ_1 and Θ_2 (two models or model-versions) is given by the difference between their weighted goodness-of-fit statistics:

$$\frac{\mathcal{L}(\Theta_1|\mathbf{y})}{\mathcal{L}(\Theta_0|\mathbf{y})} = \exp\left(-\frac{r_1^2 - r_0^2}{2}\right) \ . \tag{7}$$

Focussing on relative likelihoods removes the explicit dependence of results on n, but we are still left with two important practical issues: how many observations should be used to evaluate the model, and to what extent are they independent? In principle, all available observations could incorporated into the likelihood function, but this has undesirable

consequences in practice since all climate models fail to simulate some observable aspects of the climate system. Hence a nave incorporation of all available observations into r^2 results in comparing the relative likelihood of models whose individual likelihoods are vanishingly small. Worse, because r^2 is dominated by its largest individual terms, relative likelihoods are dominated by the difference between the simulations and those aspects of the observations that the models simulate least well.

Three approaches have been used in the literature to address this problem. In ascending order of complexity, they are: M1, metrics restricted to a subset of observable quantities that, on the basis of the evidence available, the model appears capable of simulating for at least some settings of the parameters Θ; M2, metrics in which the individual contributions to r^2 from different observation-types are renormalized by the error in the best available (or a reference) simulation of that observation-type; and M3, metrics in which the contribution of irreducible model-data discrepancies are incorporated into $\mathbf{C}$ through an explicit "discrepancy term".

In general, the choice of an M1, M2 or M3 metric will have a much greater impact on results than the choice of observations or the quality of individual models, so it is imperative to be clear which type of metric is used in any individual study. Moreover, we should not expect them to give similar results: in general, relative likelihoods based on an M1 metric will be larger (closer to unity, meaning the metric has less power in discriminating between models) than those based on an M2 or M3 metric because the M1 metric makes use of only a subset of the observations available. This does not automatically mean that the M2 or M3 metrics are preferable, because their additional power comes at the price of substantial and un-testable additional assumptions.

2.1 Option M1: restricted metrics

The convention adopted in the climate change detection and attribution literature and in the ASK approach to ensemble climate forecasting based upon it has been to assess model quality using only observable quantities that models are capable of simulating directly. In, for example, [**?**], model-simulated patterns of response to greenhouse, anthropogenic aerosol and natural (solar and volcanic) forcing were compared with observed large-scale temperature changes over the 20^{th} century using a regression analysis. In this example, the parameter vector Θ contained only three elements, being the unknown scaling factors on the responses to these three forcing agents. Principal Component Analysis was used to retain only those spatio-temporal scales of variability for which, after the best-fit Θ had been obtained, the minimum residual r^2_{min} was consistent with the expected residual due to internal climate variability (which, for large-scale temperature changes, dominates observation error), based on a standard F-test for residual consistency ([**?**]). [*Forest* et al., 2002] take a similar approach, varying only two parameters in an intermediate complexity model.

The interpretation of relative likelihoods is straightforward in this instance: for these

specific variables (large-scale temperatures) we have no reason to doubt that there is a choice of parameters Θ with which the model simulates the real-world response entirely realistically, and the likelihood of Θ_1 being that "true" set declines with $\delta r_1 = r_1^2 - r_{\min}^2$. In terms of classical statistical tests, this provides the basis for a test of the hypothesis that $r_{\min}^2$ would be this much smaller than r_1^2 if Θ_1 is in fact the "true" parameter-set.

Despite the attraction of being firmly grounded in classical linear regression and hypothesis testing, the metrics used in ASK and [*Forest* et al., 2002] are open to criticism. First, they make very limited use of the observations available, since relatively few observable quantities satisfy this condition of being statistically indistinguishable from the best-fitting available climate model simulations. Second, large-scale temperature changes are generally not the most impact-relevant aspects of a climate forecast. Applying relative likelihoods based on large-scale temperature changes to forecast changes in other variables requires an assumption that the model-simulated relationship between large-scale temperatures and these other variables is correct.

It should be noted that this second criticism does not only apply to metrics restricted to large-scale temperature changes: in general, relative likelihoods based on more complex metrics will be dominated by model-data differences in a small number of observable variables and hence require an assumption that models that simulate the observations realistically in these variables are also more likely to be realistic in other respects, although the use of an explicit discrepancy term can alleviate this problem. Perhaps the principle disadvantage, if it is one, of restricted metrics is that they are sufficiently simple that all such assumptions are out in the open.

2.2 Option M2: renormalized metrics

If more observable quantities are included in the definition of the r^2 goodness-of-fit statistic than the best-fitting models are capable of simulating (for example, by including small-scale temperature changes, or variables other than temperature that models simulate less well), then relative likelihoods tend to be dominated by these poorly simulated quantities. While this is clearly undesirable, there may still be information to be extracted from relative goodness-of-fit in these quantities: for example, the best models may be capable of simulating them realistically but they are excluded from a restricted metric simply because we lack of an adequate representation of expected model-data differences in these quantities.

A simple approach to incorporating more observations into the r^2 statistic than would be allowed under a restricted metric is simply to renormalize model-data differences in subsets of the observable quantities (putting temperatures in one subset, for example, and precipitation in another) by the average error in either the best-fit or some reference model. This means that equal weight is given, by construction, to relative errors in different subsets of the observations. This approach, used by [**?**], allows more observations to be used but lacks a clear methodological justification. We should note that the classical

statistics literature warns firmly against the use of observed residuals to renormalize expected model-data discrepancies in this way, so this approach should be regarded at best as an *ad hoc* method to be used until a more complete understanding of expected model-data differences is available.

2.3 Option M3: explicit discrepancy terms

The most sophisticated approach to incorporating a wide variety of observations into measures of model quality is the "discrepancy term" used by [**?**, **?**, **?**]. This approach incorporates all sources of model-data differences into the covariance matrix $\mathbf{C}$, including a representation of "irreducible" errors that are common to all members of the ensemble. Hence, rather than excluding observable quantities that the best-fitting models are unable to simulate or simply renormalizing model-data differences to downweight these terms, the discrepancy term attempts to include the source of these irreducible differences into $\mathbf{C}$. The result is to inflate the expected covariance in observables that the models are known to simulate poorly, which has the desirable effect of reducing the weight given these quantities in the overall measure of goodness-of-fit.

There is nothing inherently Bayesian about the use of an explicit discrepancy term: likelihoods can still be calculated on individual models using the augmented covariance matrix $\mathbf{C}$ in the weighted goodness-of-fit statistic and then applied to whatever sampling method is preferred. The justification of the discrepancy term is, however, generally framed in Bayesian terms. If we assume that we begin with an ensemble of simulations based on models drawn at random from a representative set of possible models containing the hypothetical "reified" model (i.e. a model that is not necessarily perfect, but cannot be further improved upon), and assume the differences between these simulations arise from a combination of random, parametric and structural uncertainties, then the inclusion of a discrepancy term in $\mathbf{C}$ arises naturally as a representation of structural uncertainty.

Specification of the discrepancy term presents a challenge in practice. Purely subjective methods such as expert elicitation are generally impractical for a high-dimensional covariance matrix. Methods retain, however, a strong subjective element, in that it is clearly desirable to encapsulate expert knowledge such as "all models simulate cloud height poorly, so we should minimise the weight given to errors in cloud height weight in the goodness-of-fit statistic." Such knowledge, however, is primarily of use in checking whether the discrepancy term has been adequately specified: a more formal method is required to specify it in the first place.

To date, the approach taken to estimating the discrepancy term has been to use the statistics of an independent ensemble. For example, in deriving a discrepancy term for the analysis of a perturbed-physics ensemble, [**?**] use the statistics of the multi-model ensemble available through the CMIP-3 experiment. [**?**] show that this approach is justified subject to rather limited assumptions about the properties of this second ensemble.

One assumption that is required, however, known as "second-order exchangeability", is that errors are equally probable in any two members of the multi-model ensemble. This is problematic, since in any analysis based on an ensemble-of-opportunity it is generally expected that some models in the ensemble will be substantially more realistic (through higher resolution, more advanced representation of physical processes and so on) than others. In practice, therefore, the set of second-order-exchangeable models of similar expected quality is likely to be rather small (there are typically only two or three "state-of-the-art" models available and running the relevant simulations at any given time).

Use of a multi-model ensemble to estimate the discrepancy term has intuitively attractive consequences: in particular, it incorporates information about model disagreement into the analysis, allowing less weight to be given to model-observation disagreement in variables on which model disagree among themselves. The discrepancy term is also used to allow explicitly for uncertainty in the forecast arising from errors common to all members of the perturbed-physics ensemble. Suppose, for example, an ASK-type analysis assumes, or all members of the perturbed-physics ensemble predict, a particular relationship between forecast large-scale temperature change and forecast rainfall change in a particular region. And suppose a much broader range of relationships emerges from a multi-model ensemble. In this case, it is clearly desirable to incorporate this additional uncertainty into forecast rainfall changes based on the ASK analysis or perturbed physics ensemble.

It is worth emphasising that explicit discrepancy terms play two roles in an ensemble climate forecast: one is allowing for structural uncertainty in the simulation of observable quantities $\mathbf{x}_\mathrm{o}$ that are used to constrain the forecast, while the second is allowing for structural uncertainty in the forecast $\mathbf{x}_\mathrm{f}$ itself. Although they are generally justified together, these roles are not necessarily inseparable. The use of an explicit discrepancy term to allow for structural uncertainty in $\mathbf{x}_\mathrm{f}$ is perhaps less controversial since there is no obvious alternative, short of simply refusing to issue a forecast for quantities that are not related consistently to observable quantities across both perturbed-physics and multi-model ensembles. While attractive from the perspective of methodological purity, this is unlikely to be workable in practice.

There is an alternative to the use of an explicit discrepancy term to represent structural uncertainty in $\mathbf{x}_\mathrm{o}$, which is simply to restrict metrics to observable quantities that our best models are capable of simulating. Since, for these quantities, a well-specified discrepancy term would be small, the impact of including additional observables for which a substantial discrepancy term is required should always be to reduce likelihoods relative to the best-fit model, increasing the power of the goodness-of-fit statistic to distinguish between models. The price, however, is the use of quantities that none of our models can simulate adequately, for reasons that may be completely unknown, in constraining the ensemble. Whether or not this is desirable is clearly open to debate, but it is evidently inconsistent with standard practice in much of the climate research literature to date.

3 Sampling in perturbed-physics and multi-model ensembles

Independent of the method used to assign a likelihood, or any kind of quality measure, to individual members of the ensemble, uncertainty analysis of climate forecasts also requires a method of generating the ensemble to be weighted in the first place. On a practical level, the method used to generate the ensemble is largely independent of the method used to weight individual members: each of the three metrics described above could be used in conjunction with each of the approaches to sampling described below. In general, the theoretical justification of certain metrics has typically been associated with particular approaches to sampling (M3 is normally associated with S2, for example), but the theoretical constraints are sufficiently weak that an equally coherent justification could be given for any other combination. Hence we feel it is useful to distinguish sampling approaches from model metrics.

3.1 Option S0: Ensembles-of-opportunity

The most widely-used approach to the treatment of uncertainty in climate forecasts is the ensemble-of-opportunity, typified by model intercomparison studies in which simulations from multiple modelling groups are contributed to a central repository and the spread of the ensemble is interpreted as a measure of forecast uncertainty. In general, some kind of model quality threshold is used, at least informally, to determine which models are included in the ensemble, so the ensemble-of-opportunity approach could in principle be combined with any of the three metrics described above. In practice, however, the majority of studies that use formal metrics of model quality also use a more systematic approach to sampling design.

The ensemble-of-opportunity approach has been criticised for producing forecast spreads that are potentially misleadingly narrow if all modelling groups are individually aiming to produce a best-fit model [**?**]. Conversely, however, since it has been demonstrated that it is possible to generate a very broad range of behaviour by varying parameters in models [**?**], ensembles-of-opportunity might in future produce a misleadingly wide range of uncertainty unless formal methods are used to constrain them.

As the size of ensembles-of-opportunity increases, and particularly when results have to be presented from ensembles of varying sizes, a case may be made for interpreting the ensemble as a frequency distribution, and presenting percentiles of the distribution rather than simple maximum-minimum ranges. This approach becomes problematic, however, as soon as these percentiles begin to be interpreted in probabilistic terms, particularly when some kind of model quality threshold has been used to determine which models are included in the ensemble in the first place.

3.2 Option S1: Range-over-threshold approaches

The simplest generalisation of the ensemble-of-opportunity is simply to give forecast ranges spanned by models that satisfy some formal criterion of goodness-of-fit to observations. This is the approach traditionally taken in the detection and attribution literature, and it produces classical confidence intervals, not formal probability statements. In essence, the objective is to generate an ensemble of models with a very broad range of behaviour (in detection and attribution, for example, scaling factors are imposed that span the full real number line) and then select the subset that fit the data as well or better than would be expected in, say, 90% of cases due to known sources of model-data difference.

Although results from this kind of approach have been presented in terms of probability density functions (e.g. [*Stott & Kettleborough*, 2002]), this requires some further assumptions discussed under option S3 below. The simplest interpretation of range-over-threshold approaches is in terms of confidence intervals: if all the forecasts from models that fit the data better than the $P = 0.1$ threshold lie within a certain range, and we can assume with 90% confidence that a hypothetical reified model will fit the data at least this well, then it is reasonable to assume at this confidence level that the forecast of the reified model will lie within this range.

It is an open question whether confidence intervals provide an adequate basis for the presentation of uncertainty in climate forecasting. Some applications, such as probabilistic risk assessment, require a full probability density function of future climate, but in many practical situations decision support simply requires a plausible range of outcomes, for which a classical confidence interval may well be adequate.

The advantage of range-over-threshold approaches is transparency and testability: the hypothesis that no model can be generated that yields a forecast outside a given range while simultaneously satisfying a given criterion of goodness-of-fit to observations is clearly testable and does not depend on how models or model-versions were sampled in the first place, provided the initial ensemble is broad enough and densely sampled enough to span the range consistent with relevant observations.

3.3 Option S2: Bayesian weighted ensembles

The simplest approach to generating an explicit probabilistic climate forecast is the Bayesian weighted ensemble. Under this approach, an ensemble is generated systematically by varying underdetermined inputs into the models, including model parameters but potentially also varying model structure as well, using a subjective process such as expert elicitation to assign distributions to these inputs. Individual members of the ensemble are then weighted by their likelihood with respect to observations and a posterior

distribution for forecast quantities of interest derived using Bayes theorem:

$$\Pr(\mathbf{x}_\mathrm{f}|\mathbf{y}) = \frac{\Pr(\mathbf{x}_\mathrm{o} = \mathbf{y}|\Theta)\Pr(\Theta)}{\Pr(\mathbf{x}_\mathrm{o} = \mathbf{y})} \quad . \tag{8}$$

The problem, which has been extensively documented in the literature, is that when the constraints provided by the observations are weak (meaning the likelihood function $\Pr(\mathbf{x}_\mathrm{o} = \mathbf{y}|\Theta)$ is only weakly dependent on Θ), results can be highly sensitive to the prior specification of parameters $\Pr(\Theta)$. For example, [**?**] noted that different prior specifications which had all been used in the literature resulted in a range of estimates of the upper bound on climate sensitivity spanning a factor of three or more.

One response, favoured by conventional Bayesians, is to argue that certain priors reflect investigators' beliefs better than others, and to explore sensitivity to results over "reasonable" choices of prior ([**?**, **?**]. The problem is determining what is deemed reasonable, particularly when a prior has to be specified over a model parameter, such as a diffusivity, whose physical interpretation may itself be ambiguous. To date, debates over the relative reasonableness of different prior assumptions have all taken place after the studies have been done which establish their impact on posterior forecast distributions, making it difficult to separate views about the reasonableness of different priors from the reasonableness of different posteriors. Clearly there is a danger, in this situation, of priors being tuned, perhaps subconsciously, to give an expected forecast distribution, making the whole ensemble forecasting system nothing more than an expensive way of validating the investigators' pre-conceptions.

In favour of the conventional Bayesian approach, and in contrast to the other three approaches considered here, the posterior distribution $\Pr(\mathbf{x}_\mathrm{f}|\mathbf{y})$ has an unambiguous probabilistic interpretation: it represents the investigators' degrees of belief regarding the relative probability of different forecast outcomes in the light of these observations. The problem is that Bayesian methods of presenting uncertainty are relatively uncommon in climate science outside of climate forecasting: in the analysis of past and present climate, confidence intervals are much more frequently used in the presentation of uncertainty.

3.4 Option S3: 'Objective' Bayesian approaches

One option for combining the testability and reproducibility of range-over-threshold approaches with the probabilistic interpretation of the conventional Bayesian approach is to use 'objective', or rule-based, priors to specify parameter distributions. For example, [*Allen* et al., 2000, *Stott & Kettleborough*, 2002, **?**, **?**] sample parameters to give a uniform prior predictive distribution in the quantities used to constrain the forecast. When the constraints are approximately Gaussian, as is the case in the examples considered, this is very close to the use of a Jeffreys Prior [**?**, **?**]. It has the advantage that the posterior model-simulated distribution of the quantities used to constrain the forecast is, by construction, identical to their input distributions, which is clearly desirable if we

wish the ensemble to reflect the information provided by these constraints as faithfully as possible. Bayesian purists, however, object that such uniform distributions do not reflect the actual prior beliefs of the investigators, raising questions as to what probability distribution functions obtained in this way actually mean: they do not reflect posterior beliefs regarding the odds on different outcomes, but in a situation where repeated experiments are impossible, it they cannot be interpreted in frequentist terms either.

In practice, objective Bayesian approaches may be much more widely used than is generally acknowledged, if the impact of the prior on observable quantities is, perhaps subconsciously, being taken into account in deciding what constitutes a "reasonable" prior in conventional Bayesian analyses. It is interesting that the priors recommended as reasonable for climate sensitivity in [**?**] turn out to have very similar properties over the relevant range to the uniform prior in transient climate response (TCR) recommended by [**?**] on the grounds that TCR was closer to linear in both observable and forecast quantities under most realistic scenarios than climate sensitivity. [**?**] did not make the link to a Jeffreys Prior, but it is implicit in the procedure they adopt.

References

[*Allen & Tett*, 1999] Allen, M. R., & Tett, S. F. B. Checking internal consistency in optimal fingerprinting. *Climate Dynamics*, **15**, 419–434. 1999.

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

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

[*Hasselmann*, 1993] Hasselmann, K. Optimal fingerprints for the detection of time dependent climate change. *Journal of Climate*, **6**, 1957–1971. 1993.

[*Hasselmann*, 1997] Hasselmann, K. On multifingerprint detection and attribution of anthropogenic climate change. *Climate Dynamics*, **13**, 601–611. 1997.

[*Hegerl* et al., 1996] Hegerl, G. C., von Storch, H., Hasselmann, K., Santer, B. D., Cubasch, U., & Jones, P. D. Detecting greenhouse gas-induced climate change with an optimal fingerprint method. *Journal of Climate*, **9**, 2281–2306. 1996.

[*Leroy*, 1998] Leroy, S. Detecting climate signals, some Bayesian aspects. *Journal of Climate*, **11**, 640–651. 1998.

[*Murphy* et al., 2004] Murphy, J., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*. to appear. 2004.

[*Santer* et al., 1994] Santer, B. D., Brüggemann, W., Cubasch, U., Hasselmann, K., Höck, H., Maier-Reimer, E., & Mikolajewicz, U. Signal-to-noise analysis of time-dependent greenhouse warming experiments. Part 1: pattern analysis. *Climate Dynamics*, **9**, 267–285. 1994.

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