Noname manuscript No.
(will be inserted by the editor)

Towards a typology for constrained climate model forecasts

A. Lopez, E. B. Suckling, F. E. L. Otto, A. Lorenz, D. Rowlands, M. Allen

Received: date / Accepted: date

Abstract In recent years several methodologies have been developed to combine and interpret ensembles of climate models with the aim of quantifying uncertainties in climate projections. Constrained climate model forecasts have been generated by combining various choices of metrics used to weight individual ensemble members, with diverse approaches to sampling the ensemble. The forecasts obtained are often significantly different, even when based on the same model output .

Therefore, a climate model forecast classification system can serve two roles: to provide a way for forecast producers to self-classify their forecasts; and to provide information on the methodological assumptions underlying the forecast generation and its uncertainty when forecasts are used for impacts studies.

In this review we propose a possible classification system based on choices of metrics and sampling strategies. We illustrate the impact of some of the possible choices in the uncertainty quantification of large scale projections of temperature and precipitation changes, and briefly discuss possible connections between climate forecast uncertainty quantification and decision making approaches in the climate change context.

Keywords Climate forecasts · Observational constraints · impacts studies

Myles R. Allen , F. E. L. Otto, A. Lorenz and D. Rowlands
School of Geography and the Environment, Oxford University, South Parks Rd, Oxford OX1 3NP, UK.
E B Suckling
Centre for the Analysis of Time Series, London School of Economics, Houghton Street, London WC2A 2AE, UK.
Ana Lopez
Centre for the Analysis of Time Series, London School of Economics, Houghton Street, London WC2A 2AE, UK; and AOPP, Oxford University, South Parks, Oxford OX1 3UP, UK.

1 Introduction

A number of recent reviews have examined the use of ensembles of climate models as a measure of uncertainty in climate forecasting [1]. Most conclude that interpreting the distribution of models in terms of the probability that the real world response will lie in a given interval is problematic, because it is unclear to what extent these ensembles have been designed, and can be expected, to span the range of possible behavior of the climate system [65, 30, 29, 31]. Numerous studies have attempted to use climate observations explicitly to constrain climate forecasts, in the hope of providing more robust and reproducible estimates of forecast uncertainty. However, the results of these studies are very difficult to compare, because as well as using different models, ensemble designs and observational constraints, they often rely on fundamentally different assumptions regarding the meaning of uncertainty in a climate forecast and the status of climate model output. These assumptions, which are often obscure to forecast users, have a first-order impact on estimates of forecast uncertainty.

In this review, we propose a classification system based on two broad distinctions that relate various recent studies: differences between methods used to assign a likelihood or goodness-of-fit statistic to individual ensemble members, and differences in the methods use to sample the climate models ensemble in order to generate the forecast. In section 2 we present a historical overview of the different approaches that have attempted to quantify forecast uncertainty. We then describe and categorize the metrics used to assign a likelihood to individual ensemble members in section 3, and the sampling methods to generate the forecast in section 4. Section 5 shows an example illustrating the influence of metric and sampling strategy on the forecast uncertainty quantification. Finally, in section 6 we discuss the utility of a forecast classification system for forecast users, focusing on the relationship between the approach chosen to interpret the climate ensemble information and the formal decision analysis method adopted by the decision maker.

2 Overview

Some of the earliest studies attempting to quantify uncertainty in climate forecasts emerged from the detection and attribution literature of the 1990s, notably the optimal fingerprinting approach of references [16, 17, 56, 18]. In [36, 3] the authors observed that optimal fingerprinting could be cast as a linear regression problem in which it is assumed that general circulation models (GCMs) simulate the space-time patterns of the climate response to various external drivers correctly, and observations are used to estimate the magnitude of that response. It was argued that while it is acceptable to assume that spatio-temporal patterns of response are independent of the response amplitude for large-scale surface temperature changes, this is not valid in general, in particular for changes in atmospheric circulation or precipitation, or for abrupt changes in forcing over the period of interest [2].

[1] We use the term 'forecast' as an estimation of future events, including raw and post processed climate model output, which in the case of decadal or longer time scales is conditioned on future emission scenarios .

Optimal fingerprinting can be thought of as equivalent to generating a large "pseudo-ensemble" simply by taking the mean space-time 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 the response magnitude. The goodness-of-fit between individual members of this pseudo-ensemble are then evaluated with some kind of 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 weights or metric [1, 20]. For example, the range of warming attributable to anthropogenic greenhouse gas increases over the past 50 years, evaluated 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 Assessments [21, 22].

Applying the same scaling factors to model-simulated responses to future forcing provides a method for deriving confidence intervals on future climate change [2, 62, 63], this has been referred to as the ASK (Allen-Stott-Kettleborough) approach. The crucial assumption (which is also implicit in attribution studies) is that fractional errors in model-simulated responses persist over time [2], 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 under certain forcing scenarios. A similar approach, but comparing simple or intermediate-complexity models directly with observations was taken in [9, 10, 32, 39], hereafter FKM (Forest-Knutti-Meinshausen). Other authors, such as [15, 68], also used this approach of varying parameters in simple climate models to generate uncertainty ranges or distributions for climate forecasts, but we highlight FKM since they make a point of systematic comparison of model simulations with observations. An advantage of FKM is simplicity: it is straightforward to generate large ensembles with simple and intermediate-complexity models, varying parameters to generate a broad range of behavior and then filter these by comparison with observations. The disadvantage is that direct comparison of the output of this class of models with observations is problematic, since their representation of, for example, land and ocean is inevitably idealized, making it ambiguous what observations they should be compared against (although similar issues can also be raised with GCMs).

Both ASK and FKM can provide ranges of uncertainty in forecast climate that, for variables that are poorly constrained by observations, may be much wider than the range of available GCM simulations in a multi-model ensemble such as CMIP3 or CMIP5. This was clearly an advantage when very few models were available, and will continue to be necessary as long as the spread of simulations in multi-model ensembles is thought to underestimate the full range of uncertainty. These methods therefore provide a complementary approach to more recent methods of probabilistic forecasting such as weighted multi-model ensembles [66], or perturbed-physics ensembles generated by varying model parameters using expert subjective assessments of their uncertainty [42].

Using Bayesian methods as in [66], in [42, 41, 59] the perturbed physics ensembles were weighted by their goodness-of-fit to observations, generating distributions that have an explicit probabilistic interpretation as the degree of belief in the relative probability of different outcomes in the light of the evidence available. This is arguably the simplest approach to uncertainty analysis of GCM-based climate forecasts, and the most natural for non-climate-modelers: given a complex model containing uncertain parameters, specify distributions for all these parameters, sample them to generate an ensemble and constrain with observations. Difficulties in the implementation of this approach arise because many of the parameters to which climate forecasts are particularly sensitive do not correspond to any observable and only really mean something in the context of a particular model or parametrization, and hence cannot be assigned a standard error as an observable quantity might be.

The sheer number of under-determined parameters in climate models also makes it impossible to ensure that, for a given model structure, all important uncertainties have actually been sampled. This is illustrated graphically in [53, 54], where nominally similar parameters were varied over nominally similar ranges in two GCMs obtaining a very broad distribution of responses in one case and a relatively narrow one in the second case.

We conclude this overview by noting that these different approaches have used very different underlying statistical philosophies. 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. Early implementations of FKM were also frequentist in character, while recent implementations [55,39] have used more explicitly Bayesian approaches, exploring sensitivities to prior distributions but still generally avoiding any claim to accurate representation of actual subjective prior beliefs. In contrast, the studies in references [42,41,59] have generally aimed to provide credible intervals, or Bayesian posterior probabilities - ranges within which the forecast quantity of interest is expected to lie given both the prior expectations of the investigators and the constraints of the observations. [2]

These different approaches should only be expected to give similar results if the observations provide a very strong constraint on the forecast quantity of interest, which is typically not the case in the long-term climate forecasting problem. 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 modeling groups are aiming to simulate observations as well as possible), these conditions are not satisfied, so ranges provided by the different approaches are not directly comparable. It would be helpful to forecast users to be clearer about which approach is being used in the presentation of uncertainty in any particular study. In what follows we suggest a classification scheme that could be used to facilitate this task.

3 Metrics of individual model quality

All but the simplest approaches to sampling a range of uncertainty on a climate model forecast require some measure of the quality of individual climate models or model-versions. In general, this can be characterized as a distance measure, often expressed as a weighted sum squared difference between a model simulation $\mathbf{x}$,

[2] The distinction between confidence intervals and credible intervals is best summarised thus: if a forecast quantity lies outside a 90% confidence interval, then an event has occurred that was estimated at the time of the forecast to have a less than 10% probability of occurrence. If the forecast quantity lies outside a 90% credible interval, then the forecast quantity is found to have a value inconsistent (at the 10% level) with our expectations at the time the forecast was made.

which may be the mean of an initial-condition ensemble, and the corresponding set of observations $\mathbf{y}$:

$$r^2 = (\mathbf{y} - \mathbf{x}_o)^{\mathbf{T}} \mathbf{C}^{-1} (\mathbf{y} - \mathbf{x}_o) \ , \tag{1}$$

where $\mathbf{C}$ is a measure of the expected difference between model and observations due to processes that can be treated as random. It typically represents 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 $\mathbf{x}_o$ and $\mathbf{C}$ are determined by a set of parameters Θ, the model-observations deviance can be expressed as a likelihood:

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

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 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 present, but for the sake of unity, we will assume that structural uncertainty can in principle be parameterised.

In a Bayesian analysis, the likelihood $\mathcal{L}(\Theta|\mathbf{y})$ is simply proportional to the probability density function of obtaining a simulation $\mathbf{x}_o$ in the vicinity of $\mathbf{y}$ given the parameters Θ, $\Pr(\mathbf{x}_o = \mathbf{y}|\mathbf{\Theta})$. Clearly, this tends to become progressively smaller the higher the dimension of $\mathbf{y}$ simply because the probability of the simulation "hitting the target" falls off the higher the dimension of the 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 the relative likelihood:

$$\frac{\mathcal{L}(\Theta_1|\mathbf{y})}{\mathcal{L}(\Theta_0|\mathbf{y})} = \exp\left(-\frac{r_1^2 - r_0^2}{2}\right) \ , \tag{3}$$

where Θ_1 and Θ_0 are two sets of parameters (two models or model-versions). 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 be incorporated into the likelihood function, but this has undesirable consequences in practice since all climate models fail to simulate many observable aspects of the climate system. Hence a naïve 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 [42]. This will result in poorly constrained model variables having a disproportionately greater impact in the weighting.

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". There is of course another possibility (M0), which is not to use any metric at all and consider all the ensemble members as equally likely.

In general, the choice of a metric will have a 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 generally un-testable additional assumptions.

3.1 Option M1: restricted metrics

The convention adopted in ensemble climate forecasting based upon climate change detection and attribution approaches, has been to assess model quality using only observable quantities that models are capable of simulating directly. For example in reference [61], model-simulated space-time 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, Θ contained 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 [3]. In [10] only three parameters are varied in an intermediate complexity model, also using principle component analysis to focus on the large-scale response.

The interpretation of relative likelihoods here is straightforward: for these specific variables (large-scale temperatures changes) the assumption is that there is a choice of parameters Θ with which the model simulates the real-world warming response realistically, and the likelihood of Θ_1 being that "true" set declines with $\delta r_1 = r_1^2 - r^2_{\text{min}}$. In terms of classical statistical tests, this provides the basis for a test of the hypothesis that r^2_{min} 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 FKM are open to criticism. First, they make very limited use of the observations available, since few observable quantities satisfy the 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 constrain forecast changes in other variables requires the strong assumption that the model-simulated relationship between large-scale temperatures and these other variables is correct. Alternatively, it could be argued that for impacts studies, relative likelihoods should include a comparison with observations relevant to the particular application, such as rainfall and evaporation for hydrological impacts. However, choosing to constrain the uncertainty range using observables that models cannot skillfully simulate will simply discard many models, potentially resulting in a reduced uncertainty range as a result of the inadequacy of the models and not a genuine reduction of the uncertainty.

It should be noted that the 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. The key advantage of restricted metrics, however, is that they are sufficiently simple that all such assumptions are out in the open.

3.2 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 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 in [45,60], allows more observations to be used but lacks a clear methodological justification, so it 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.

3.3 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 in [41,59,49,7] to estimate likelihoods of individual models in a perturbed physics ensemble. Rather than excluding observable quantities that the best-fitting models are unable to simulate, as with M1, or simply renormalizing model-data differences to down-weight these terms as in M2, the discrepancy term in M3 attempts to include all the sources of model-data differences into **C**, including a representation of "irreducible" errors that are common to all members of the ensemble. The result is to inflate the expected covariance in observables that the models are known to simulate poorly, which has the effect of reducing the weight given to these quantities in the overall measure of goodness-of-fit.

Specification of the discrepancy term presents a challenge in practice. 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, [59] use the CMIP3 experiment. [12] 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. However, it is generally expected that some models will be substantially more realistic than others (through higher resolution, more advanced representation of physical processes and so on). In practice, therefore, the set of second-order- exchangeable models of similar expected quality is likely to be rather small. Use of a multi-model ensemble to estimate the discrepancy term 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.

It is worth emphasizing that explicit discrepancy terms play two roles in an ensemble climate forecast: one is allowing for structural uncertainty in the simulation of the observable quantities that are used to constrain the forecast, while the second is allowing for structural uncertainty in the forecast itself. Although they are generally justified together, these roles are not necessarily inseparable.

4 Sampling in perturbed physics and multi-model ensembles

Independent of the climate model ensemble generation technique and the method used to assign a quality measure to individual members of the ensemble , there are different sampling methods to generate the climate forecast. In general, the theoretical justification of certain metrics has typically been associated with particular approaches to ensemble sampling, but the theoretical constraints are sufficiently weak that an equally coherent justification could be given for any combination. Hence it is useful to distinguish ensemble sampling approaches from model metrics.

4.1 Option S0: Unrestricted ensembles

The most widely-used approach for the treatment of uncertainty in climate forecasts is the multi-model ensemble or ensemble-of-opportunity, typified by model intercomparison studies in which simulations from multiple modeling groups, are contributed to a central repository and the spread of the ensemble is interpreted as a measure of forecast uncertainty. These include for instance the CMIP3 [38] and CMIP5 [64] global modeling intercomparison projects, as well as regional downscaling experiments such as CORDEX [25]. Ensembles-of-opportunity 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 ensemble sampling. The ensemble-of-opportunity approach has been criticised for producing forecast spreads that are potentially misleadingly narrow if all modelling groups are individually tuning their models aiming to produce a best-fit model [26].

An alternative approach to treat uncertainty in the forecasts consist of constructing a perturbed physics ensemble. In this case, as shown in [60], it is possible to generate a very broad range of behaviors. Therefore, this type of unrestricted ensembles might in principle produce a misleadingly wide range of uncertainty, unless formal methods are used to constrain the ensemble.

A specific issue with the interpretation of ensembles-of-opportunity is whether the models in such an ensemble represent approximations to the real world each subject to independent errors, or whether the real world should be regarded as interchangeable with a member of the ensemble [5]. This has practical implications, since the "truth-plus-error" interpretation implies that as the ensemble size increases, the ensemble mean should converge steadily closer to the truth, as the impact of independent errors cancel out, whereas the "exchangeable" interpretation implies no such convergence. On climate timescales model errors cannot be assumed to be mutually independent, so the mean of a large ensemble is no more likely to be closer to the truth than the mean of a small ensemble simply by virtue of the ensemble size. Hence, with some exceptions [66], analyses of ensembles-of-opportunity have tended to treat them as if ensemble members were interchangeable with the real world (e.g. [46]).

The two interpretations discussed above assume that climate models are adequate representations of the Earth climate system, and model inadequacies represent small perturbations around the true system. There is of course a third possibility, whereby the climate models structural uncertainties are severe enough to invalidate their use as climate forecasting tools, particularly at spatial and temporal scales relevant for impacts studies (see for example [48]). In that case, uncertainties estimated using any of the approaches discussed here do not necessarily represent the true uncertainty range, and forecast users should consider this possibility particularly when using the forecast for decision support.

4.2 Option S1: Range-over-threshold approaches

The simplest generalisation of measuring forecast uncertainty as the spread of the climate model ensemble, be it a multi-model or a perturbed physics ensemble, is to provide 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, given the ensemble of models with a very broad range of behaviour, a subset is selected 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.

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. This is correct provided that the initial ensemble is broad enough and densely sampled enough to span the range consistent with relevant observations.

4.3 Option S2: Bayesian approaches

The simplest approach to generating an explicit probabilistic climate forecast is the Bayesian weighted ensemble. Under this approach, individual members of the climate model ensemble, usually a perturbed physics ensemble but can include sampling of model structure uncertainty as well, are weighted by their likelihood with respect to observations, and a posterior distribution for forecast quantities of interest is derived using Bayes theorem. In this framework, the posterior distribution provides credible intervals, and represents the investigators' degrees of belief regarding the relative probability of different forecast outcomes in the light of these observations.

A limitation of this approach is that, when the constraints provided by the observations are weak (meaning that the likelihood function is only weakly dependent on the parameters), results can be highly sensitive to the prior specification of parameters. For example, [11] 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, is to argue that certain priors reflect investigators' beliefs better than others, and to explore sensitivity to results over "reasonable" choices of prior [58,4]. Determining what is deemed reasonable, however, is not straightforward, particularly when a prior has to be specified over a model parameter, such as a diffusivity, whose physical interpretation may itself be ambiguous.

An 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, [2,62,13,11] sample parameters to give approximately uniform prior predictive distributions in the quantities used to constrain the forecast. When the constraints are approximately Gaussian and independent, as is the case in the examples considered, this is very close to the use of a Jeffreys prior [23,24,50,37] to specify parameter distributions.

5 Implications for uncertainty quantification

In this section, we illustrate the effect of the choice of sampling-metric combinations on the quantification of the forecast uncertainty. For our illustration we use the model data and an example of a metric M1 described in detail in [51]. In this work the metric evaluates the distance between the simulated and observed large scale spatio-temporal temperature anomalies, over the period 1961-2010.

Figure 1 shows how the unceratinty range of projections of temperature and precipitation changes depends on the combination of M1 with two sampling strategies (S0 and S1) , when applied to two climate model ensembles: an ensemble of opportunity (CMIP3) and a perturbed physics ensemble (climateprecition.net). Projections for precipitation vs. temperature changes are shown for the global mean and three sub-continental regions. As expected, M1 based only on large scale spatio-temporal temperature anomalies, works well at constraining the global mean warming projections since the range of temperatures (horizontal axis) spanned by the unrestricted climateprediction.net ensemble (colored crosses, S0-M0) is much wider than the range-over-threshold ensemble (grey diamonds, S1-M1) in the top right panel of the figure. However, using a metric based on the models' ability to simulate large scale spatio-temporal temperature patterns is not equally effective in all regions. For instance, for the climateprediction.net ensemble, the range-over-threshold for Southern Asia and Western North America temperature projections is better constrained than for Northern Europe. This can be explained by the fact that for climateprediction.net models there is a strong relationship between global and regional warming projections for the first two regions while the relationship is weaker for the third region(not shown). In the case of the CMIP3 unrestricted ensemble (circles, S0-M0), the uncertainty range is not reduced when applying the metric M1 (solid grey circles, S1-M1), possibly because this multi model ensemble is, by construction, tuned with the observations used to build the metric M1 [52], so imposing this constraint does not add new information to costraint the uncertainty range.

The figure also illustrates that the metric M1 is not very effective at constraining projections for precipitation changes (vertical axis) for the regions shown. In these modeling experiments, the model projections for large scale temperature changes used to compute M1 do not have a strong relationship with simulated changes in precipitation for those particular regions(not shown), therefore model performance in large-scale temperature changes does not provide useful information to constrain the uncertainty in projections on precipitation changes.

In the context of impacts studies, it is important to remark that a metric that evaluates warming patterns does not provide information about absolute errors and biases in models. In other words, even though it might make sense to constrain the range of uncertainty in projections using an observable that can be adequately simulated by at least some of the models (warming patterns in this case), that does not imply that models which pass the test according to one metric are realistic at simulating more relevant quantities for impacts studies, such as absolute values of variables (as opposed to their anomalies).

6 Discussion and Conclusion

The simple example above illustrates clearly that the choice of ensemble sampling strategy and goodness-of-fit metric has a strong influence on the forecast uncertainty range. As it is well known, the uncertainty in projections for unrestricted ensembles is significantly different depending on the modeling strategy (CMIP3 vs climateprediction.net). When observations are used to constrain uncertainty ranges, the result depends not only on which observations (and what temporal and spatial scales) are used to construct the metric, but also on the relationships between that information and the forecasted variables. In our example, comparison with observations is restricted to large scale temperature warming patterns which in principle the best models in the ensemble can simulate realistically. Even in this case we see that the M1 metric used to quantify model quality can constrain the uncertainty range in global mean temperature changes, but it fails to constrain it in precipitation changes.

The proliferation of approaches to uncertainty analysis of climate forecasts is clearly unsatisfactory from the perspective of forecasts users. When confronted with a new forecast with a nominally smaller range of uncertainty than some alternative, it would take considerable insight to work out if the difference results from arbitrary changes in metric, or ensemble sampling, or from new information that reduces the uncertainty in the forecast. While it is undesirable to impose a common approach, it may be useful for studies to attempt some form of self-classification. The typology proposed in this review is intended to provide a starting point to this end, with the expectation that this classification will evolve in time to incorporate perhaps new approaches applicable to other types of ensembles generated using for instance pattern scaling [40, 67] or stochastic parametrisations [44].

Such a classification could serve as a guide for those attempting to use climate forecasts for impact studies and formal decision analysis in the climate change context. Especially for the latter, the assumptions about the decision criterion employed in the analysis are naturally related to the assumptions underlying the generation of the climate forecast. Scenario analysis, robust control ([35]), or info-gap (e.g. [14]) frameworks do not rely on probabilistic information or even ranges, but focus on the impacts of decision options and system response under a range of possible futures. However, the applicability of these types of analysis to future decisions rests on a sufficiently comprehensive coverage of the space of possible future climate states.

Climate ensembles providing a range of possible futures can be utilised in decision analysis using the MaxiMin (pessimistic), MaxiMax (optimistic) or Hurwicz (mixture) criteria [33], which only rely on information about the worst and/or best possible outcomes. The expected utility decision criterion (e.g. see [8]) is widely used in cost-benefit (e.g. [43]), cost-risk (e.g. [57]), and cost-efficiency (e.g. [19]) analyses in the climate change context as the current standard of normative decision theory. It requires information about the climate forecasts in the form of probability density functions (pdfs), and naturally relates to Bayesian ensemble sampling approaches. However, among many other shortcomings, the expected utility criterion is unable to represent a situation in which the decision maker is ambiguous about the exact pdfs representing (climate) uncertainty (see e.g. [34]). One possible solution to this shortcoming is the use of imprecise probabilities (e.g. [27], [28]), where climate information would be given not as a single pdf, but as a set of possible pdfs.

We close this discussion by remarking that, when considering climate forecasts for impacts studies, it is important to keep in mind that, as discussed in section 4 the possible range of climate changes might not be fully explored if the analysis relies solely on climate models' projections. Changes other that the ones currently projected by climate models are plausible, particularly at impacts relevant spatial scales. Therefore decision makers should use a variety of scenarios for their planning, and not restrict their analysis exclusively to model projected ranges of uncertainties [47,6].

Acknowledgements FELO and EBS acknowledge support from the NERC EQUIP project (NE/H003479/1). AL acknowledges support from the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re.

References

1. Allen, M.R.: Liability for climate change. Nature **421**, 891–892 (2003)
2. Allen, M.R., Stott, P.A., Mitchell, J.F.B., Schnur, R., Delworth, T.L.: Quantifying the uncertainty in forecasts of anthropogenic climate change. Nature **407**, 617–620 (2000)
3. Allen, M.R., Tett, S.F.B.: Checking for model consistency in optimal fingerprinting. Climate Dynamics **15**, 419–434 (1999)
4. Annan, J.D., Hargreaves, J.C.: Using multiple observationally-based constraints to estimate climate sensitivity. Geophysical Research Letters **33**(L06704) (2006)
5. Annan, J.D., Hargreaves, J.C.: Reliability of the CMIP3 ensemble. Geophysical Research Letters **37**(L02703) (2010)
6. Brown, C., Wilby, R.L.: An alternate approach to assessing climate risks. EOS, Transactions American Geophysical Union **93**(41), 401 (2012)
7. Collins, M., Chandler, R.E., Cox, P.M., Huthnance, J.M., Rougier, J., Stephenson, D.B.: Quantifying future climate change. Nature Climate Change **2**(6), 403–409 (2012)
8. Fishburn, P.C.: The foundations of expected utility. Theory & Decision Library **31**, 176 (1982)
9. Forest, C.E., Allen, M.R., Stone, P.H., Sokolov, A.P.: Constraining uncertainties in climate models using climate change detection techniques. Geophysical Research Letters **27**(4), 569–572 (2000)
10. 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**(5552), 113–117 (2002)
11. Frame, D.J., Booth, B.B.B., Kettleborough, J.A., Stainforth, D.A., Gregory, J.M., Collins, M., Allen, M.R.: Constraining climate forecasts: The role of prior assumptions. Geophysical Research Letters **32**(L09702) (2005)
12. Goldstein, M., House, L., Rougier, J.: Assessing model discrepancy using a multi-model ensemble. MUCM Technical Report **08/07** (2008)
13. Gregory, J., RJ, S., SCB, R.: An Observationally Based Estimate of the Climate Sensitivity. Journal of Climate **15**, 3117–3121 (2002)
14. Hall, J.W., Lempert, R.J., Keller, K., Hackbarth, A., Mijere, C., McInerney, D.J.: Robust Climate Policies Under Uncertainty: A Comparison of Robust Decision Making and Info-Gap Methods. Risk Analysis **32**(10), 1657–1672 (2012)
15. Hansen, J., Rossow, W., Carlson, B., Lacis, A., Travis, L., Genio, A.D., Fung, I., Cairns, B., Mishchenko, M., Sato, M.: Low-cost long-term monitoring of global climate forcings and feedbacks. Climatic Change **31**, 247–271 (1995)
16. Hasselmann, K.: Optimal fingerprints for the detection of time-dependent climate change. Journal of Climate **6**, 1957–1971 (1993)
17. Hasselmann, K.: On multifingerprint detection and attribution of anthropogenic climate change. Climate Dynamics **13**, 601–611 (1997)
18. 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)
19. Held, H., Kriegler, E., Lessmann, K., Edenhofer, O.: Efficient climate policies under technology and climate uncertainty. Energy Economics **31**(0), S50–S61 (2009)
20. Huntingford, C., Stott, P.A., Allen, M.R., Lambert, F.H.: Incorporating model uncertainty into attribution of observed temperature change. Geophysical Research Letters **33**(L05710) (2006)
21. IPCC: Climate change 2001: The Scientific Basis. . Cambridge University Press (2001)

22. IPCC: Climate change 2007: The Physical Science Basis. . Cambridge University Press (2007)
23. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London Series A **186**(1007), 453–461 (1946)
24. Jewson, S., Rowlands, D., Allen, M.: A new method for making objective probabilistic climate forecasts from numerical climate models based on Jeffreys' Prior. arXiv **physics.ao-ph**(0908.4207v1) (2009)
25. Jones, C.F.G., Asrar, G.: The Coordinated Regional Downscaling Experiment (CORDEX). An international downscaling link to CMIP5. Tech. rep., International CLIVAR Project Office, Southampton,UK (2011)
26. Kiehl, J.: Twentieth century cimate model response and climate sensitivity. Geophysical Research Letters **34**(L22710) (2007)
27. Klibanoff, P., Marinacci, M., Mukerji, S.: A Smooth Model of Decision Making under Ambiguity. Econometrica **73**, 1849–1892 (2005)
28. Klibanoff, P., Marinacci, M., Mukerji, S.: Recursive smooth ambiguity preferences. Journal of Economic Theory **144**(3), 930–976 (2009)
29. Knutti, R.: The end of model democracy. Climatic Change **102**, 395–404 (2010)
30. Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G.A.: Challenges in combining projections from multiple climate models. Journal of Climate **23**, 2739–2758 (2010)
31. Knutti, R., Masson, D., Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there. Geophysical Research Letters **40**, 1194–1199 (2013)
32. Knutti, R., Stocker, T.F., Joos, F., Plattner, G.K.: Constraints on radiative forcing and future climate change from observations and climate model ensembles. Nature **416**, 719–723 (2002)
33. Lange, A.: Climate change and the irreversibility effect combining expected utility and MaxiMin. Environmental & Resource Economics **25**(4), 417–434 (2003)
34. Lange, A., Treich, N.: Uncertainty, learning and ambiguity in economic models on climate policy: some classical results and new directions. Climatic Change **89**(1-2), 7–21 (2008)
35. Lempert, R.J.S.W.P.S.C.B.: Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Tech. rep., RAND Corporation, Santa Monica, CA: (2003)
36. Leroy, S.: Detecting climate signals, some Bayesian aspects. Journal of Climate **11**, 640–651 (1998)
37. Lewis, N.: Noninformative prior distributions for observationally based objective estimates of climate sensitivity PDFs. unpublished (2013)
38. Meehl, G.A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J.F.B., Stouffer, R.J., Taylor, K.E.: The WCRP CMIP3 multi-model dataset: A new era in climate change research. Bulletin of the American Meteorological Society **88**, 1383–1394 (2007)
39. Meinshausen, M., Meinshausen, N., Hare, W., Raper, S.C.B., Frieler, K., Knutti, R., Frame, D.J., Allen, M.R.: Greenhouse gas emission targets for limiting global warming to 2C. Nature **458**(08017), 1158–1163 (2009)
40. Mitchell, T.D.: Pattern scaling. An examination of the accuracy of the technique for describing future climates. Clim. Change **60**, 217–242 (2003)
41. Murphy, J.M., Booth, B.B.B., Collins, M., Harris, G.R., Sexton, D.M.H., Webb, M.J.: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philosophical Transactions of the Royal Society A **365**, 1993–2028 (2007)
42. Murphy, J.M., 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 **430**(02771), 768–772 (2004)
43. Nordhaus, W.D.: A Question of Balance, 1 edn. Yale University Press (2008)
44. Palmer, T.N.: Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. Q.J.R.Meteorol. Soc. **138**, 841–861 (2012)
45. Piani, C., Frame, D.J., Stainforth, D.A., Allen, M.R.: Constraints on climate change from a multi-thousand member ensemble of simulations. Geophysical Research Letters **32**(L23825,) (2005)
46. Räisänen, J., Palmer, T.N.: A probability and decision-model analysis of a multimodel ensemble of climate change simulations. Journal of Climatic **14**, 2312–3226 (2001)
47. Risbey, J.S.: Sensitivities of Water Supply Planning Decisions to Streamflow and Climate Scenario Uncertainties. Water Policy **1**(3), 321–340 (1998)

48. Risbey, J.S., O'Kane, T.: Sources of knowledge and ignorance in climate research. Climatic Change **108**(4), 755–773 (2011)
49. Rougier, J.C.: Probabilistic inference for future climate using an ensemble of climate model evaluations. Climatic Change **81**, 247–264 (2007)
50. Rowlands, D., Jewson, S., Meinshausen, N., Frame, D., Allen, M.: Quantifying uncertainty in climate projections using Jeffrey's prior. unpublished (2013)
51. Rowlands, D.J., Frame, D.J., Ackerley, D., Aina, T., Booth, B.B.B., Christensen, C., Collins, M., Faull, N., Forest, C.E., Grandey, B.S., Gryspeerdt, E., Highwood, E.J., Ingram, W.J., Knight, S., Lopez, A., Massey, N., McNamara, F., Meinshausen, N., Piani, C., Rosier, S.M., Sanderson, B.M., Smith, L.A., Stone, D.A., Thurston, M., Yamazaki, K., Yamazaki, Y.H., Allen, M.R.: Broad range of 2050 warming from an observationally constrained large climate model ensemble. Nature Geoscience **5**, 256–260 (2012)
52. Sanderson, B.M., Knutti, R.: On the interpretation of constrained climate model ensembles. Geophysical Research Letters **39**, L16,708 (2012)
53. Sanderson, B.M., Knutti, R., Aina, T., Christensen, C., Faull, N., Frame, D.J., Ingram, W.J., Piani, C., Stainforth, D.A., Stone, D.A., Allen, M.R.: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. Journal of Climate **21**, 2384–2400 (2008)
54. Sanderson, B.M., Shell, K.M., Ingram, W.: Climate feedbacks determined using radioactive kernels in a multi-thousand member ensemble of AOGCMs. Climate Dynamics **35**(7) (2010)
55. Sansó, B., Forest, C., Zantedeschi, D.: Inferring climate system properties from a computer model. Bayesian Analysis **3**(1), 57–62 (2008)
56. 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)
57. Schmidt, M.G., Lorenz, A., Held, H., Kriegler, E.: Climate targets under uncertainty: challenges and remedies. Climatic Change **104**(3-4), 783–791 (2011)
58. Sexton, D.M.H., Murphy, J.M.: Multivariate probabilistic projections using imperfect climate models. Part II: robustness of methodological choices and consequences for climate sensitivity. Climate Dynamics **38**(11-12), 2543–2558 (2012)
59. Sexton, D.M.H., Murphy, J.M., Collins, M., Webb, M.J.: Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. Climate Dynamics **38**(11-12), 2513–2542 (2012)
60. Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., Piani, C., Sexton, D., Smith, L.A., Spicer, R.A., Thorpe, A.J., Allen, M.R.: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature **433**(03301), 403–406 (2005)
61. Stott, P.A., Jones, G.S., Lowe, J.A., Thorne, P., Durman, C., Johns, T.C., Thelen, J.C.: Transient climate simulations with the HadGEM1 climate model: causes of past warming and future climate change. Journal of Climate **19**(12), 2763–2782 (2006)
62. Stott, P.A., Kettleborough, J.A.: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. Nature **416**(6882), 723–626 (2002)
63. Stott, P.A., Mitchell, J.F.B., Allen, M.R., Delworth, T.L., Gregory, J.M., Meehl, G.A., Santer, B.D.: Observational constraints on past attributable warming and predictions of future global warming. Journal of Climate **19**(13) (2006)
64. Taylor, K., Stouffer, R., Meehl, G.: An Overview of CMIP5 and the experiment design. Bull. Amer. Meteor. Soc. **93**, 485–498 (2012)
65. Tebaldi, C., Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. R. Soc. A **365**(2053-2075) (2008)
66. Tebaldi, C., Mearns, L.O., Nychka, D., Smith, R.W.: Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. Geophysical Research Letters **31**(L24213) (2004)
67. Watterson, I., Whetton, P.: Probabilistic projections of regional temperature and precipitation extending from observed time series. Climatic Change **119**, 677–691 (2013). DOI 10.1007/s10584-013-0755-y
68. Wigley, T.M.L., Raper, S.C.B.: Interpretation of high projections for global-mean warming. Science **293**(5529), 451–454 (2001)

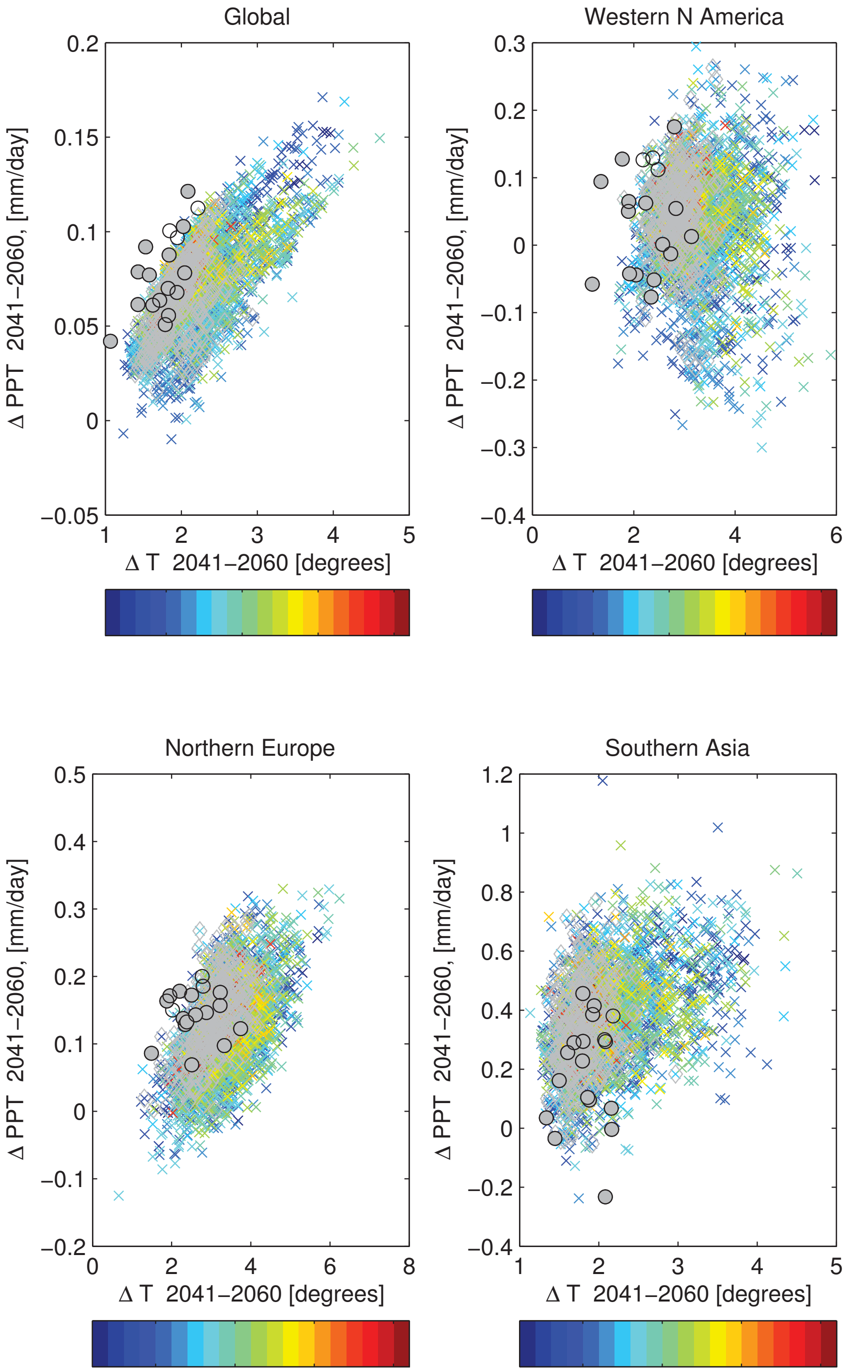

Fig. 1 Illustration of metric-sampling combinations. Projections for precipitation vs. temperature changes are shown for the global mean and three sub-continental regions as indicated in the panels. CMIP3 ensemble of opportunity: S0-M0 (all circles), S1-M1 (solid grey circles). Climateprediction.net perturbed physics ensemble: S0-M0 (colored crosses), S1-M1 (grey diamonds). The metric M1 evaluates the goodness of fit of large scale spatial and temporal temperature anomalies over 1961-2010. For the PPE the color of the crosses corresponds to the number of parameters that have been perturbed in each model version, from red indicating no perturbed parameter to dark blue indicating 19 parameters perturbed simultaneously.