Special Issue Article

Environmetrics

Received: 19 October 2011, Revised: 20 April 2012, Accepted: 05 June 2012, Published online in Wiley Online Library:

(wileyonlinelibrary.com) DOI: 10.1002/env.2153

Statistical problems in the probabilistic prediction of climate change

David B. Stephenson[a], Matthew Collins[a], Jonathan C. Rougier[b] and Richard E. Chandler[c]

Future climate change projections are constructed from simulated numerical output from a small set of global climate models—samples of opportunity known as *multi-model ensembles*. Climate models do not produce probabilities, nor are they perfect representations of the real climate, and there are complex inter-relationships due to shared model features. This creates interesting statistical challenges for making inference about the real climate.

These issues were the focus of discussions at an Isaac Newton Institute workshop on probabilistic prediction of climate change held at the University of Exeter on 20–23 September 2010. This article presents a summary of the issues discussed between the statisticians, mathematicians, and climate scientists present at the workshop. In addition, we also report the discussion that took place on how to define the concept of climate. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: climate; ensemble; prediction; inference

1. INTRODUCTION

In a presentation on the unsolved problems in mathematics on 8 August 1900 at the International Congress of Mathematicians in Paris, Professor David Hilbert made this remark (Hilbert, 1902):

> *The deep significance of certain problems for the advance of mathematical science in general and the important role which they play in the work of the individual investigator are not to be denied. As long as a branch of science offers an abundance of problems, so long is it alive; a lack of problems foreshadows extinction or the cessation of independent development. Just as every human undertaking pursues certain objects, so also mathematical research requires its problems. It is by the solution of problems that the investigator tests the temper of his steel; he finds new methods and new outlooks, and gains a wider and freer horizon.*

By this standard, climate science is very much alive and is raising many interesting and unsolved statistical problems. Climate science has made outstanding progress in the past 50 years with the development of ever more complex and more detailed numerical models of the climate system (Randall *et al*., 2007; Chandler *et al*., 2010). Climate model simulations are numerical solutions to a large set of nonlinear first order differential equations subject to boundary conditions or ‘forcings’ (greenhouse gases, solar radiation, aerosols, etc.) that control the subsequent evolution of the system. It is now common practice to sample uncertainty in climate predictions by considering ensembles of simulations from one or more climate models. Ensembles are created either by perturbing a climate model’s initial conditions (initial condition ensembles) or physical parameters (perturbed physics ensembles), or by taking model outputs from a set of different climate models (multi-model ensembles; MME). For example, the forthcoming Intergovernmental Panel on Climate Change Fifth Assessment Report intends to present projections based on data from an MME of around 40 climate models produced by climate modelling centres around the world as part of the coordinated CMIP5 project.[1] Typical time series from an MME are shown in Figure 1 together with corresponding observations of UK temperature.

Despite their increasing complexity and seductive realism, it is important to remember that climate models are *not* the real world. Climate models are numerical approximations to fluid dynamical equations forced by parameterisations of physical and unresolved sub-grid scale processes. Climate models are inadequate in a rich diversity of ways, but it is the hope that these physically motivated models can still inform us about various aspects of future observable climate. A major challenge is how we should use climate models to construct credible

* *Correspondence to: David B. Stephenson, Exeter Climate Systems, Mathematics Research Institute, University of Exeter, Exeter, U.K. E-mail d.b.stephenson@exeter.ac.uk*

a *Exeter Climate Systems, Mathematics Research Institute, University of Exeter, Exeter, U.K.*

b *Department of Mathematics, University of Bristol, Bristol, U.K.*

c *Department of Statistical Science, University College London, London, U.K.*

[1] Freely available climate model data from Coupled Model Intercomparison Project Phase 5 (http://cmip-pcmdi.llnl.gov/cmip5)

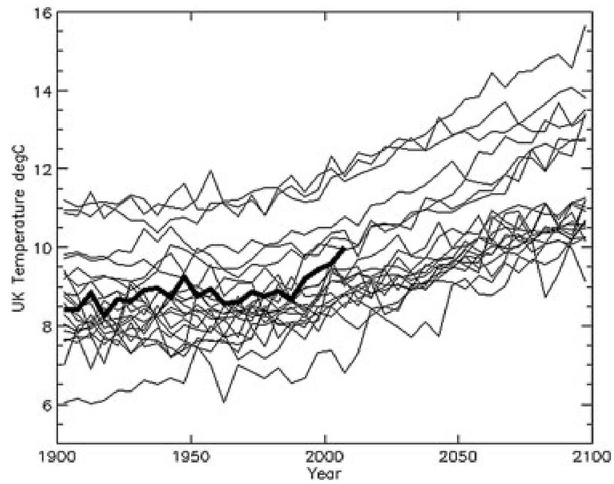

Figure 1. Example of multi-model ensemble output from 22 climate models: UK temperatures from the CMIP3 experiment (single black lines) and historical observations (thick black line). The statistical challenge is how to infer from the ensemble what might happen to the observation time series in the future.

probabilistic forecasts of future climate. Because climate models do not themselves produce probabilities, an additional level of explicit probabilistic modelling is necessary to achieve this.

Various methods have been used to obtain future climate projections from ensembles of climate model output, for example, the simplest pragmatic approaches of calculating equally weighted means of the ensemble members, or unequally weighted means of the ensemble members based on descriptive model performance metrics. Alternatively, one can use statistical frameworks such as regression of future changes on past model statistics (e.g. Bracegirdle and Stephenson, 2012), or more complex hierarchical models (Buser *et al*., 2009; Knutti *et al*., 2010a; Leith and Chandler, 2010; Collins *et al*., 2012; references therein). However, there is little agreement on what is the most reliable and robust methodology for making probabilistic predictions of real-world climate.

To help address this important and pressing challenge, we proposed probabilistic climate prediction as one of the two major themes for a residential research programme at the Isaac Newton Institute for Mathematical Sciences in Cambridge, entitled 'Mathematical and Statistical Approaches to Climate Modelling and Prediction' (Chandler *et al*., 2010). The programme brought together more than 150 climate scientists and mathematicians/statisticians over the period August–December 2010. As part of the interdisciplinary programme, we organised a workshop on probabilistic climate prediction at the University of Exeter from 20 to 23 September 2010. The main aims of the workshop were to bring climate and statistical experts together to discuss, identify, and formulate the big and potentially solvable problems in the mathematics of probabilistic climate prediction. More details of format, talks, and participants are provided in the Appendix.

The purpose of this article is to summarise the interesting discussions that took place at the workshop in order to open up these issues to wider participation from statisticians. We hope that the article is also useful to climate scientists wishing to know more about the more fundamental issues involved in interpreting ensembles and making probabilistic climate predictions—it is important to be aware of the statistical modelling issues. One of the first stages in a problem is to recognise that a problem exists, and so it is our hope that this article will unveil some of the more challenging problems that are still in serious need of solutions. It should be noted that as a reflection of the diverse ideas that emerged at the Exeter workshop, this article is neither a consensus statement nor is it a comprehensive account of all the interesting statistical problems that arise in climate science.

2. FUNDAMENTAL PROBLEMS

Three major problems were proposed by the organisers at the start of the workshop:

1. How should MMEs be interpreted statistically?
2. How might design of experiments to be used to construct more informative MMEs?
3. How can we best account for climate model discrepancy and inadequacy?

The participants agreed that these were problems worthy of discussion, and they also proposed a fourth problem: How to statistically define and interpret the concept of climate. The remainder of this section summarises the breakout and plenary discussions on these four problems starting with the fundamental problem of how to define and interpret climate.

2.1. What is climate?

Climate is a surprisingly beguiling concept. Operationally, climate is defined as a sample statistic of weather (e.g. the World Meteorological Organisation's definition of climate as the average weather over the period 1961–1990). However, such a sample statistic definition is rather limited and a much deeper statistical and scientific insight is possible by considering climate to be the process that generates weather.

To focus on climate prediction, we discussed how one might define 'the climate in Exeter on 1 Jan 2070'. Workshop participants agreed that this local concept made sense and that it is synonymous with 'the probability distribution of weather in Exeter on 1 Jan 2070'.

The 'weather' that occurs in Exeter on 1 January 2070 is considered to be random a draw from this probability distribution. Note that climate here is more than just the expectation of weather—it is the whole probability distribution including extremes. Moving beyond this particular location in space and time, climate can be more generally modelled as a space–time stochastic process whose sample paths are the observable weather. Whereas weather may be directly observed, climate is a process that has to be inferred by using a probability model.

Were we in Exeter on 1 January 2070, we would experience only a single realisation drawn from the distribution. Replication, required for estimating the distribution, is obtainable by making statistical judgements about similarity and independence/exchangeability of weather over neighbouring times. For example, one might assume that climate is stationary, and hence that weather is identically distributed on each of the 31 days, in January 2070. More generally, climate in Exeter on a certain date could be defined using weather observations in nearby spatial locations close to the date of interest. The key point here is that to estimate the 'probability distribution of weather' from weather observations requires a statistical model.

Alternatively, one may attempt to construct the probability distribution of weather from an ensemble of simulations from a climate model. For an ergodic[2] climate model with time-invariant boundary conditions, one may estimate the distribution of weather by taking samples from a single long simulation: the 'distribution' representing climate is the invariant measure (state space attractor) for the system. However, climate forcings are not generally time-invariant, and so in practice we must be more subtle. To find the probability distribution of weather in Exeter on 1 January 2070, we could create replicates by considering an ensemble of simulations out to 2070 starting at random initial times during a period when boundary conditions may be assumed to be time-invariant (e.g. pre-industrial). Such an approach also assumes that the climate model is a perfect representation of the real climate system and that the climate model is ergodic. Because climate models are imperfect representations of the real climate, calibration of the climate model runs is generally required, for example, bias adjustment of the means. Moreover, *real* future forcings are of course unknown and so the climate estimated using this approach is conditional on the forcing 'scenarios' that are used.

2.2. How should multi-model ensembles be interpreted statistically?

Climate model output contains complex dependencies. For any climate model, there is a complex relationship between the output variables (indexed by location, time, and type) due to the regularities that arise from basic physical principles and from equations of state. There are also complex dependencies between output from different climate models due to the presence of common code modules, common resolution, common coupling schemes, and so on. Climate models often evolve from earlier versions and so one might consider a complex evolutionary tree showing the ancestry of each climate model. Some climate models are close to one another in this tree, whereas others are rather unrelated with few common components (Masson and Knutti, 2011). Some species of models are no longer 'viable' and these are superseded by more advanced models. Typically, unviable models are those with numerical approximations that have been rendered obsolete, such as flux corrections in the ocean, or low-resolution models that now look grossly unrealistic in comparison with higher resolution versions. Model evolution is also partially determined by the ability of models to reproduce historic observations—models with poor fits are usually modified or even replaced by model developers.

A rigorous analysis of the information that an MME contains about actual climate should ideally account for everything we know about all of the models and their inter-relationships. This is not feasible, and so it is necessary to make simplifying assumptions in order to make the problem tractable. All climate predictions based on MMEs make such simplifications, either implicitly (e.g. ensemble mean predictions) or explicitly by formulating statistical modelling assumptions. It should be noted that different assumptions lead to different climate change predictions and this can be a substantial additional source of indeterminacy that also needs to be assessed and reported (e.g. Ho *et al*., 2012).

The simplest most widely used approach for making climate change predictions is to take (weighted) sample means over the ensemble runs, for example, calculate a point estimate of the mean climate change response by taking the difference between the mean of all future simulations and the mean of all present day simulations. The choice of weights for calculating the ensemble means is not obvious and various heuristic approaches are used. For example, climate change studies have used equal weights for all simulations, or alternatively, for each climate model participating in a recent modelling experiment (e.g. CMIP5 runs), or 'metrics' (descriptive statistics) based on the ability of each climate model to simulate various aspects of past climate. Although this weighted mean approach can provide a point estimate of the mean response, it makes no judgements on dependencies between models or how climate models might be related to the real world and so it cannot quantify the uncertainty in the climate change response. The normalised weights should not be naively interpreted as probabilities assigned to each climate model being equal to the real climate because the members of an MME do not represent a partition of sample space, for example, the complement of a particular climate model, or the intersection of two climate models is not defined.

To create credible probabilistic interpretations of climate predictions, assumptions need to be made explicit and transparent by proposing credible statistical frameworks for modelling the joint distribution of the MME and the real climate. The development and testing of such frameworks are an essential component of climate modelling that requires more concerted collaboration between statisticians and climate scientists. To create such frameworks, one needs to develop a probabilistic interpretation of the MME and a probabilistic interpretation of how real climate might be related to the MME. Because we only have one realisation of real climate, the relationship of real climate to the MME requires a subjective Bayesian interpretation of probability. However, both frequentist and Bayesian interpretations are possible for interpreting the climate model outputs in the MME. To avoid confusion, we should clarify that these interpretations are relevant for how one formulates a statistical model of the MME, rather than how inference might ultimately be performed from the resulting statistical model. In what follows, we discuss probabilistic interpretations of MMEs based on discussions at the workshop and our own subsequent

[2]A dynamical system whose trajectories cover all points in state space so that expectations over state space can be well approximated by time averages over a trajectory.

thoughts. It should be noted that alternatives to probability theory for quantifying uncertainty have been proposed, however, these have not yet been applied to high-dimensional climate models and so will not be covered here but can be found in the Professor Arthur Dempster's lecture cited in the Appendix.

2.2.1. *Frequentist interpretation of the MME*

Random sampling experiments play a crucial role in our inferences about the real world, particularly in medicine, agriculture, and engineering. The random sampling concept considers the available MME (or a quality-controlled subset of it) as a sample from a notional population of 'potential' climate models, and hence allows us to make inferences about this population. Such an interpretation has the advantage that many non-technical people (and many climate scientists) feel reassured if they can imagine a probability statement in relation to a very large number of climate simulations produced by a very large number of simulators under a large number of plausible parameter and input perturbations. Such a 'super-ensemble' of many climate worlds would represent in some sense the sum total of our possible knowledge. To interpret probabilities in terms of some underlying population, one must be clear about what exactly the population of climate models is and how our MME has been sampled from it. Without this, we are deluding ourselves in trying to attribute any conceptual interpretation to the probabilities.

So what is the 'ideal' population of viable climate models? There are two stages in the development of any climate model. The first stage of model development involves selecting and assembling 'components' from a large range of possible options (e.g. numerical schemes, processes to include, parameterisations, etc.). However, the selection is weighted in favour of particular components or combinations of components, based either upon community consensus or upon opinions regarding (for example) the extent to which particular schemes represent the detailed processes that are believed to be operating in the real climate system, or even due to computer resource constraints. Thus, the allowable combinations of components satisfy some constraints of what might be called 'structural credibility'. The second stage involves testing to see if the result is capable of reproducing aspects of past climate; this might involve some tuning or development/replacement of some components. One might then consider the population from which the MME is drawn to be the set of all possible structurally credible combinations of simulator components for which at least some of the parameter values produce 'acceptable' simulations of past climate.

How are the climate models in the MME sampled from the population of viable models? At a basic level, the sample is determined by decisions made about the specific choice of components from which to build the climate model. In climate science, these decisions are made by small groups of model developers at different modelling centres around the world and one could consider modelling this as a random process. For example, the decision to accept a model component could be modelled as a random trial, for example, tossing a 'weighted coin' with a certain probability of success. However, workshop discussions argued that this simplistic view of model development is not defensible because climate models generally evolve from previous versions rather than being assembled as new for each version. In other words, the development process is evolutionary whereby successive versions of a climate model (species) are subject to gradual change and occasional mutation, with the weakest species dying out over time. Each new species of climate model is likely to be more similar to currently existing species than if the set of viable species were sampled completely randomly each time. In addition, models are not developed in isolation and there is frequent communication between development groups. So rather than a random sample, it is likely that MMEs represent a relatively small, and probably relatively homogeneous, sample from the population of viable models. The use of climate metrics to select subsets of the MME could lead to even greater reduction in the sampled range of structural uncertainty. Because of these relationships between viable climate models, the outputs in MMEs are not independent as have been found empirically by noting that the sample variance of the MME does not decrease as the reciprocal of the ensemble size (Knutti *et al*., 2010b)

2.2.2. *Bayesian interpretation of the MME*

The Bayesian approach does not need to invoke a population of climate models or a sampling mechanism. It relies instead on judgements such as conditional independence and/or exchangeability of climate model outputs, which can be made explicit mathematically. Following Rougier (2007), denote the simulated output from m climate models by $\{X_1, X_2, \ldots, X_m\}$ and the real climate by Y, where X and Y are random vectors containing variables in both the past and future. The ensemble mean of the MME is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The distribution of real climate Y may be obtained from observations Z if one knows the distribution of measurement errors (i.e. data assimilation - see Stephenson *et al*., 2005).

In making judgements, the direction of conditioning is important. Some studies have assumed that the climate model outputs are independently normally distributed about the true climate, in other words, $X_i|Y$are independently normally distributed with mean Y for $i = 1, 2, \ldots, m$ (Furrer *et al*., 2007; Tebaldi and Knutti, 2007; Smith *et al*., 2009; and references therein). This assumption that model output is truth-plus-error implies that the variance of the MME ensemble mean, $Var\bar{X}$, exceeds the variance of the real climate $Var(Y)$ no matter how large the ensemble may be. Alternatively, other studies condition observations on climate model outputs (e.g. Kettleborough *et al*., 2007; and references therein). These approaches, often used in detection and attribution studies, assume that $Z|\bar{X}$is normally distributed with a mean $a\bar{X} + b$ i.e. they linearly regress observations on climate model output.

An alternative approach is to simplify the specification of a joint distribution over the members of the MME and actual climate using judgements of *exchangeability* (Schervish, 1995), which permit a parsimonious representation of dependence. To treat the members of a MME as exchangeable is to assert that the names of the climate models convey no information, that is, probability distributions of the climate model outputs should be invariant to permutations of the model names. Although this is clearly not the case for all climate models, it may be a reasonable approximation for appropriate subsets of the MME. For example, it may be necessary to thin out climate model versions from the same modelling centre that are too similar or climate models that are known to be very different from the majority in terms of their modelling schemes. It is less defensible to assert that the MME (or some subset) and actual climate are also jointly exchangeable (Annan and Hargreaves, 2010; Knutti *et al*., 2010a). Although this *perfect model* approach might work for highly aggregated simulator outputs (e.g. global mean temperature in 2100), most climate scientists would expect to be able to identify real climate among the ensemble members in more detailed outputs. Rougier *et al*. (2012) use the weaker constraint that real climate respects exchangeability with the MME, which is to say that no climate model is judged, *a priori*, to be more or less accurate than any other. This introduces the notion of a *systematic discrepancy* between the MME and actual climate; a similar statistical framework for climate models has been proposed by Chandler (2011). Exchangeability can be implemented within a fully probabilistic framework, or it can be implemented in a much less demanding second-order Bayes linear framework that considers only means, variances, and covariances (Goldstein, 1986; Rougier *et al*., 2012). A second-order exchangeable set of climate model outputs has the same expectations, the same variances, and the same covariance between any pair of models. These approaches capture the shared discrepancy by separately conditioning real climate and climate model outputs onto a common underlying variable.

2.3. How might design of experiments be used to construct more informative MMEs?

Uncertainty in future climate predictions arises from several sources: (a) uncertainty in future forcing related to changes in greenhouse gases and aerosol concentrations, solar activity, and so on; (b) uncertainty in climate model parameters; (c) uncertainty in initial conditions for climate simulations; and (d) uncertainty due to choice of model structure. Multi-model ensemble simulations for different scenarios aim to sample these different sources of uncertainty.

For long-range predictions many decades ahead, scenario uncertainty is clearly an important source of uncertainty. However, it was not clear how relevant initial condition perturbations are for long-lead forecasts, for example, are initial condition uncertainties important for 100-year ahead predictions of ocean circulation or ice sheets? Stochastic sub-grid scale schemes can capture some of the effects of structural uncertainty, but certainly not all (although the use of stochastic parameterisation is only just starting in climate modelling). How should one account, for example, for the limitations of the climate model that does not simulate a Gulf Stream of the correct strength or in the correct location? One possibility is to 'retro-fit' such a process by imposing fluxes after having inspected the climate model output; but such a procedure cannot incorporate any feedbacks to the rest of the simulated climate system. It is not clear whether one should correct model discrepancies by flux correction or by statistical post-processing (e.g. the use of anomalies).

Incorporating two-way feedbacks with emission scenarios is crucial: this means that scenarios should ideally be defined by coupling climate models directly to economics models. This is obviously a paradigm shift compared with current practice, which considers climate change given fixed scenarios (Cox and Stephenson, 2007). The use in CMIP5 of prescribed emissions (representative concentration pathways) rather than concentrations is a step towards allowing more interaction in the carbon cycle. Interestingly, the discussions noted that perturbed physics parameter experiments also perturb the initial conditions obtained after spin-up from a unique initial state, because it can take different amounts of time for the climate variables to stabilise under different model parameter settings. Quite what the distribution of such perturbations represents is less obvious.

The relative merits of perturbing parameters in one model and of multi-model ensembles were discussed. Perturbing parameters in one model can be useful for quantifying parametric uncertainty caused by not knowing the best choice of parameters for a particular model. Because climate models have many parameters (typically more than 30), there are potentially a very large number of computationally expensive sensitivity tests required to explore this source of uncertainty. Emulators can be of use here. However, a single model approach does not address uncertainty that could arise from different formulations of a climate model, that is, structural uncertainty. To do this, one needs to explore multi-model ensembles to ensure that a wider range of possible future outcomes are covered (the MME spread is greater in practice than in perturbed physics ensembles; Collins *et al*., 2011).

The question of how best to design ensembles to account for the computational costs of increasing resolution and increasing ensemble size was also discussed. It was noted that climate simulations are likely to be used for many purposes because of the computational and manpower costs involved. Therefore, models need to be robust because of the many questions likely to be addressed. It was agreed that a sequential design involving 'expensive' climate models and 'cheaper' statistical emulators is probably a good solution. For a particular small subset of output variables relevant to a particular application, an emulator could be used to successively reduce the parameter space covered and to guide the 'few' runs of say a high resolution climate model that are required to achieve better accuracy. Dimension reduction gives benefit in reducing effort by using simpler models in a hierarchy of models to inform subsequent or more complex models along/up a chain of development. As well as mean quantities, there is a need to develop and test climate emulators able to handle the extreme tails of distributions because of their importance for climate impacts. See Rougier and Sexton (2007) for a discussion of emulators in climate prediction. It was noted that the existence of a utility function that could be optimised would help in the efficient design of multi-model climate model experiments. A properly formulated statistical framework for MMEs could help provide such a utility function rather than have to rely on *ad hoc* design of experiments.

2.4. How can we best account for climate model discrepancy and inadequacy?

Climate models do not perfectly represent processes in the real world and this leads to discrepancies (biases/errors) between climate model output and actual true climate. To provide some focus, the discussion focussed around the simple linear scaling equation, $Y = aX + b$, often used in climate predictions (e.g. Kettleborough *et al*., 2007). The following questions were discussed:

1. How might the model biases change in the future?
2. How can model bias best be accounted for?
3. How can we test if calibration methods are well formulated?

Question 1 arose because it was noted that climate model biases can depend heavily on processes that are likely to change in the future. For example, surface temperature biases can be very sensitive to the presence or absence of snow cover in the model (Sexton *et al*., 2012). Such dependence of biases on the underlying state is now being increasingly recognised by the climate science community especially in the Arctic predictions, which are strongly dependent on the presence or absence of sea ice (Bracegirdle and Stephenson, 2012; and references therein). Sometimes biases can also reduce in the future, for example, if the soil dries out in response to increasing temperatures. These examples have been found by explicitly looking for them and it would be useful if such state dependency on underlying processes could be included in future statistical frameworks for MMEs. Statistical frameworks need to be developed that are capable of including knowledge about how biases depend on physical processes.

There was also an interesting discussion about when should one discard a model rather than bias correct—for example, if a process is very poorly represented in a model and leads to a large additive bias. Furthermore, it should be noted that validation of individual components and their associated errors in a climate model is a necessary but not sufficient condition for confidence in the model—new and large errors can often emerge once the components are fully coupled.

Question 2 resulted from the observation that current bias correction procedures give different results depending on how one defines the 'anomalies' (Buser *et al*., 2009; Ho *et al*., 2012). This is an important additional source of uncertainty in climate prediction that arises from statistical indeterminacy due to not having a complete framework for representing the relationship between model output and reality. The issue of calibration of climate model output is often swept under the rug by the use of anomalies in the climate science community and could benefit from the development of transparent statistical models rather than *ad hoc* removal of sample means. It was noted that climate modelling centres spend much time generating climate model output but relatively little time on building statistical frameworks that can account for the biases in the climate model output. In a more holistic approach, the climate model and the statistical framework are both crucial parts of a probabilistic climate prediction system and so both deserve recognition. This is an area where there is scope for much greater collaboration between statisticians and climate scientists.

Question 3 on calibration methods deserves much more scrutiny by statisticians and climate scientists. Various approaches are used to bias correct model output (Buser *et al*., 2009; Ho *et al*., 2012) yet there is not much critical review published on the strengths and weaknesses of the various approaches that are used. Rather than advocacy of individual approaches, a more balanced critical comparison of the various approaches is required in which the assumptions in each approach are explicitly stated, and where possible, challenged. This could help eliminate unsound frameworks and help quantify the uncertainty that arises from the calibration of climate predictions.

3. CONCLUSION

To make reliable probabilistic predictions of future climate, we need probability models based on credible and defensible assumptions.

If climate is considered to be a probability distribution of all observable features of weather, then one needs well-specified probability models to define climate from the single realisation of weather provided by nature. In an increasingly non-stationary climate forced by accelerating rates of global warming, the time average over a fixed period provides only an incomplete description perhaps better suited to a more stationary pre-industrial world.

We also need probability models (or more generally a statistical framework) to make inference about future observable climate based on ensembles of simulated data from numerical climate models. Such frameworks involve making simplifying, yet transparent and defensible, assumptions about ensembles of climate models and their relationship to the real world. These frameworks not only have to model dependency between different climate models, but also need to account for model discrepancies (biases) and how these might evolve in the future. Development and testing of such frameworks are a pressing interdisciplinary challenge in statistical and climate science. Furthermore, it is likely that good statistical frameworks might be useful in the sequential design of climate model experiments, which take increasingly large amounts of time on the world's fastest supercomputers.

To solve these pressing problems, there needs to be much better recognition of the importance of probability models in climate science and a more integrated view of climate modelling whereby climate prediction involves the fusion of numerical climate models and statistical models. This challenge will require greater collaboration and understanding between climate scientists and statisticians. Details of project areas are given in the final programme report (http://www.newton.ac.uk/reports/1011/clpdraft.pdf).

ACKNOWLEDGEMENTS

We wish to thank the Isaac Newton Institute (http://www.newton.ac.uk/programmes/CLP/) and the Willis Research Network (http://www.willisresearchnetwork.com) for kindly supporting this important workshop. We also gratefully acknowledge all the participants, and especially the rapporteurs, at the workshop for their invaluable contributions and the many stimulating discussions. Finally, we wish to thank Phil Sansom and the reviewers of this manuscript for their useful comments. Any errors or omissions in the reporting of the workshop are the sole responsibility of the authors.

REFERENCES

Annan JD, Hargreaves JC. 2010. Reliability of the CMIP3 ensemble. Geophysical Research Letters 37: L02703. DOI: 10.1029/2009GL041994.

Bracegirdle TJ, Stephenson DB. 2012. More precise predictions of future polar winter warming estimated by multi-model ensemble regression. Climate Dynamics, published online, DOI: 10.1007/s00382-012-1330-3.

Buser CM, Künsh HR, Lüthi D, Wild M, Schär C. 2009. Bayesian multi-model projection of climate: bias assumptions and interannual variability. Climate Dynamics 33: 849–868.

Chandler RE. 2011. Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Research Report No. 311*. Department of Statistical Science, University College London. (Available from: http://www.ucl.ac.uk/statistics/research/reports)
Chandler RE, Rougier J, Collins M. 2010. Climate change: making certain what the uncertainties are. Significance 7(1): 9–12.
Collins M, Booth BBB, Bhaskaran B, Harris G, Murphy JM, Sexton DMH, Webb MJ. 2011. A comparison of perturbed physics and multi-model ensembles: model errors, feedbacks and forcings. Climate Dynamics 36: 1737–1766.
Collins M, Chandler RE, Cox PM, Huthnance J, Rougier JC, Stephenson DB. 2012. Perspective on quantifying future climate change. Nature Climate Change 2: 403–409.
Cox PM, Stephenson DB. 2007. A changing climate for prediction. Science 317: 207–208.
Furrer R, Sain SR, Nychka D, Meehl GA. 2007. Multivariate Bayesian analysis of atmosphere–ocean general circulation models. Environmental and Ecological Statistics 14(3): 249–266.
Goldstein M. 1986. Exchangeable belief structures. Journal of the American Statistical Association 81: 971–976.
Hilbert D. 1902. Mathematical problems. Bulletin of the American Mathematical Society 8: 437–479.
Ho CK, Stephenson DB, Collins M, Ferro CAT, Brown SJ. 2012. Calibration strategies: a source of additional uncertainty in climate change projections. Bulletin of the American Meteorological Society 93: 21–26. DOI: http://dx.doi.org/10.1175/2011BAMS3110.1
Kettleborough JA, Booth BBB, Stott PA, Allen MR. 2007. Estimates of uncertainty in prediction of global mean surface temperature. Journal of Climate 20(5): 843–855.
Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA. 2010a. Challenges in combining projections from multiple climate models. Journal of Climate 23: 2739–2758.
Knutti R, Abramowitz G, Collins M, Eyring V, Gleckler P, Hewitson B, Mearns L. 2010b. Good practice guidance paper on assessing and combining multi model climate projections. In Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections, Stocker TF, Qin D, Plattner G-K, Tignor M, Midgley P (eds). IPCC Working Group I Technical Support Unit, University of Bern: Bern, Switzerland.
Leith NA, Chandler RE. 2010. A framework for interpreting climate model outputs. Journal of the Royal Statistical Society, Series C 59(2): 279–296.
Masson D, Knutti R. 2011. Climate model genealogy. Geophysical Research Letters 38: L08703.
Randall DA, Wood RA, Bony S, Colman R, *et al*. 2007. Climate models and their evaluation. In Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). Cambridge UP: Cambridge, United Kingdom and New York, NY, USA.
Rougier JC. 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. Climatic Change 81: 247–264.
Rougier JC, Sexton DMH. 2007. Inference in ensemble experiments. Philosophical Transactions of the Royal Society A 365: 2133–2143.
Rougier JC, Goldstein M, House L. 2012. Second-order exchangeability analysis for multi-model ensembles. *Statistics Research Report 12:01*, University of Bristol. www.maths.bris.ac.uk/research/stats/reports/2012/1201.pdf, in submission.
Schervish MJ. 1995. Theory of Statistics. Springer: New York. Corrected second printing, 1997.
Sexton DMH, Murphy JM, Collins M, Webb ME. 2012. Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. Climate Dynamics 38(11–12): 2513–2542.
Smith RL, Tebaldi C, Nychka D, Mearns LO. 2009. Bayesian modeling of uncertainty in ensembles of climate models. Journal of the American Statistical Association 104(485): 97–116.
Stephenson DB, Balmaseda M, Coelho CAS, Doblas-Reyes FJ. 2005. Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. Tellus A 57(3): 253–264.
Tebaldi C, Knutti R. 2007. The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society A 15 (365): 2053–2075.

APPENDIX: THE EXETER WORKSHOP ON PROBABILISTIC CLIMATE PREDICTION

The 3-day workshop brought together 54 climate scientists and statisticians from around the world (Figure 2 and Table 1). The scene was set by several 45-min overview talks[3]:•

1. ‘Problems in probabilistic prediction of climate’ David Stephenson
2. ‘Model processes, errors and inadequacies’ Matthew Collins
3. ‘Methodologies for probabilistic uncertainty assessment’ Richard Chandler
4. ‘Using a perturbed physics ensemble to make probabilistic climate projections for the UK’ David Sexton
5. ‘Probabilistic use of climate catastrophe multi-models’ Gero Michel
6. ‘Notes on fundamental approaches to climate prediction’ Arthur Dempster
7. ‘Probabilistic frameworks’ Jonathan Rougier

The overview talks were interspersed with five short parallel breakout group sessions on four main themes. This interactive format was very successful at stimulating many interesting discussions and generating brainstorming in small groups.

[3]Slides and videos are available from http://www.newton.ac.uk/programmes/CLP/clpw02p.html

Figure 2. Participants at the University of Exeter workshop on Probabilistic Climate Prediction, 20–23 September 2010

Table 1. List of workshop participants. A good international mix of statisticians and climate scientists was achieved with participants coming from academia and national weather services such as the nearby Met Office based in Exeter

Name	Institution	Country
MR Allen	University of Oxford	UK
CW Anderson	University of Sheffield	UK
JD Annan	Japan Agency for Marine-Earth Science and Technology	Japan
B Booth	Met Office	UK
P Challenor	University of Southampton	UK
R Chandler	University College London	UK
M Collins	University of Exeter	UK
P Cox	University of Exeter	UK
DT Crommelin	Centrum voor Wiskunde en Informatica	Netherlands
M Crucifix	Université Catholique de Louvain	Belgium
AP Dempster	Harvard University	USA
S Dessai	University of Exeter	UK
J Duan	Illinois Institute of Technology	USA
NR Edwards	The Open University	UK
CL Farmer	University of Oxford	UK
CAT Ferro	University of Exeter	UK
K Fraedrich	Universität Hamburg	Germany
T Fricker	University of Exeter	UK
V Garreta	Centre national de la Recherche Scientifique	France
J Gatticker	Los Alamos National Laboratory	USA
M Goldstein	University of Durham	UK
HM Hanlon	University of Edinburgh	UK
JC Hargreaves	Japan Agency for Marine-Earth Science and Technology	Japan
G Harris	Met Office	UK
E Hawkins	University of Reading	UK
H Held	Potsdam Institute for Climate Impact Research	Germany
L Hermanson	University of Reading	UK
JM Huthnance	National Oceanography Centre, Liverpool	UK
IT Jolliffe	University of Exeter	UK
TE Jupp	University of Exeter	UK
RW Katz	University Corporation for Atmospheric Research	USA
K Keller	Pennsylvania State University	USA
F Kwasniok	University of Exeter	UK
L McColl	Met Office	UK
D McNeall	Met Office	UK
G Michel	Willis Group	UK
JC Rougier	University of Bristol	UK

(*Continues*)

Table 1. *(Continued)*

Name	Institution	Country
D Rowlands	University of Oxford	UK
M Semenov	Rothamsted Research	UK
D Sexton	Met Office	UK
DA Stainforth	London School of Economics	UK
DB Stephenson	University of Exeter	UK
P Stott	Met Office	UK
S Tett	University of Edinburgh	UK
J Thuburn	University of Exeter	UK
R Tokmakian	Naval Postgraduate School	USA
A Turasie	University of Exeter	UK
NM Urban	University of Princeton	USA
R Vitolo	University of Exeter	UK
NW Watkins	British Antarctic Survey	UK
RD Wilkinson	University of Nottingham	UK
PD Williams	University of Reading	UK
X-S Yang	National Physical Laboratory	UK
S Yip	University of Exeter	UK