PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

rsta.royalsocietypublishing.org

Research

Cite this article: Chandler RE. 2013 Exploiting strength, discounting weakness: combining information from multiple climate simulators. Phil Trans R Soc A 371: 20120388.
http://dx.doi.org/10.1098/rsta.2012.0388

One contribution of 13 to a Theme Issue 'Mathematics applied to the climate system'.

Subject Areas:
atmospheric science, climatology, statistics

Keywords:
empirical Bayes, ensemble of opportunity, general circulation model (GCM), multi-model ensemble, regional climate model, weighting

Author for correspondence:
Richard E. Chandler
e-mail: richard@stats.ucl.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsta.2012.0388 or via http://rsta.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

Exploiting strength, discounting weakness: combining information from multiple climate simulators

Richard E. Chandler

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

This paper presents and analyses a statistical framework for combining projections of future climate from different climate simulators. The framework recognizes explicitly that all currently available simulators are imperfect; that they do not span the full range of possible decisions on the part of the climate modelling community; and that individual simulators have strengths and weaknesses. Information from individual simulators is automatically weighted, alongside that from historical observations and from prior knowledge. The weights for a simulator depend on its internal variability, its expected consensus with other simulators, the internal variability of the real climate and the propensity of simulators collectively to deviate from reality. The framework demonstrates, moreover, that some subjective judgements are inevitable when interpreting multiple climate change projections: by clarifying precisely what those judgements are, it provides increased transparency in the ensuing analyses. Although the framework is straightforward to apply in practice by a user with some understanding of Bayesian methods, the emphasis here is on conceptual aspects illustrated with a simplified artificial example. A 'poor man's version' is also presented, which can be implemented straightforwardly in simple situations.

1. Introduction

(a) Multi-model ensembles

Projections of future climate are usually derived from climate simulators such as general circulation and regional climate models. Although these simulators

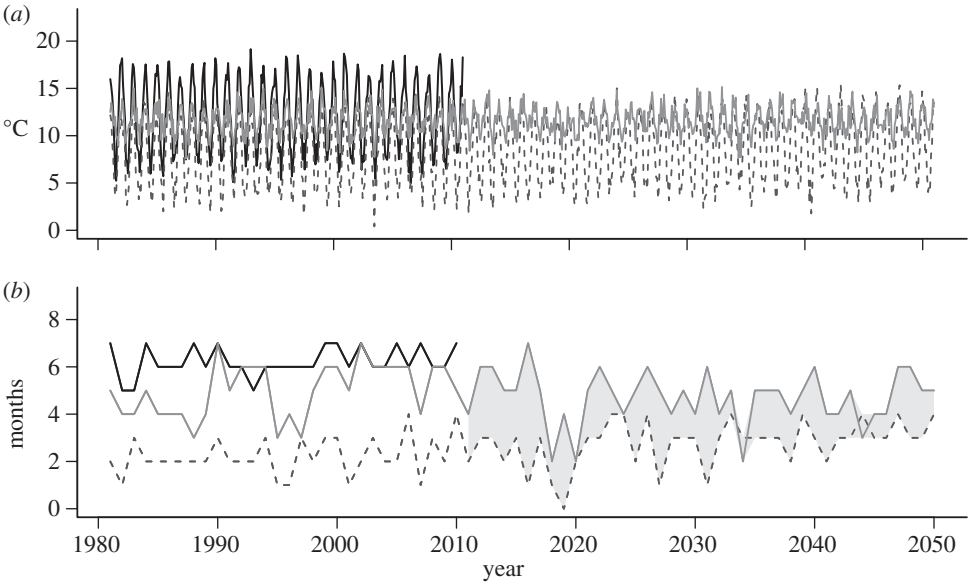

Figure 1. Hypothetical example illustrating potential use of simulator projections to assess impacts of climate change. (*a*) Monthly temperature time series from 'observations' (black solid line) and two climate simulators (black dashed line, simulator 1; grey solid line, simulator 2). Observations span the period 1980–2010; simulator runs are from 1980 to 2050. (*b*) Annual series of growing season length (defined as the number of months in a year with a mean monthly temperature in excess of 12°C) derived from each of the temperature series in (*a*). Shaded region denotes the range of attainable projections derived from any weighted average of outputs from the two simulators.

represent our best available understanding of the climate system, projections can vary appreciably between them to the extent that the choice of simulator is sometimes the dominant source of uncertainty in such projections [1]. There is increasing recognition, therefore, that the prudent user of climate change projections needs to consider information from an 'ensemble' of different simulators or 'multi-model ensemble' (MME).

The interpretation of MMEs is not straightforward, however. Tebaldi & Knutti [2] and Knutti [3] review the issues involved; a brief summary is given here. The simplest approach is to average each quantity of interest over all simulators, possibly weighting each one according to some measure of 'quality' as in the 'Reliability Ensemble Averaging' (REA) method of Giorgi & Mearns [4]; see Fowler & Ekström [5] and Christensen *et al.* [6] for examples of alternative weighting schemes. The approach has two drawbacks, however. First, it is heuristic: different weighting schemes, each based on plausible criteria, can yield different results [7,8]. Therefore, without a formal justification for the choice of weighting scheme, this choice now becomes an additional source of uncertainty [9]. Moreover, an emerging body of research suggests that an unweighted average of all simulator outputs often performs favourably by comparison with performance-based weighting schemes [6,10,11].

In many applications, simultaneous projections of several climate variables are required. However, no single simulator is uniformly better than all others [7,12]: each has strengths and weaknesses. This highlights the second drawback of simulator weighting: if projections of several different quantities are required then the rankings of the simulators, and hence the implied weights, may differ between quantities of interest. Often therefore, the usefulness of attaching scalar-valued weights to individual simulators is questionable. Figure 1 provides an artificial example to illustrate this point. Figure 1*a* shows a hypothetical monthly temperature time series that may be considered to correspond to 'observations', along with the outputs from two climate simulators (in fact, all three series have been stochastically generated, but the structure is sufficiently realistic to illustrate the main points—the code used to generate this figure, and to

carry out the analyses reported later in the paper, is provided in the electronic supplementary material). The 'observed' series ends in the year 2010, but both 'simulated' series continue until 2050. Note that the mean temperature for simulator 2 agrees reasonably well with the observations for the 1980–2010 period, but the seasonal cycle is severely dampened. By contrast, simulator 1 underestimates the mean temperature although its seasonal cycle appears reasonable. Suppose now that these simulators are to be used to inform strategies for the adaptation of agricultural practice to future climate change, and specifically to learn whether or not conditions will become more or less favourable for the cultivation of a particular crop, which requires a mean monthly mean monthly temperature of at least 12°C for growth (the precise definition is unimportant: the deliberately stylized representation is adopted so as to focus clearly on the main conceptual issues). Figure 1*b* then shows the annual time series of growing season lengths corresponding to each of the temperature series in figure 1*a*. Both of the simulator-derived series underestimate the growing season length, but for different reasons: simulator 1 because of the overall bias in its temperature projections and simulator 2 because its summers are too cool. If future projections of growing season length are to be obtained by weighting the outputs from each of the simulators, the results are necessarily constrained to lie within the shaded region in figure 1*b*—unless the possibility of negative weights is allowed, which seems unnatural without convincing formal justification. As noted by Leith & Chandler [13], in such situations it appears desirable to adopt a strategy that exploits the strengths of individual simulators while discounting their weaknesses.

Another difficult issue is how to quantify the uncertainty associated with projections from MMEs. When weights are attached to individual simulators, a simple approach is to treat these weights as defining a probability distribution as in Déqué & Somot [10] and Watterson [14], for example. Indeed, the notion of weights as probabilities is implicit in the calculation of a weighted average. However, it is far from clear that a collection of weights, derived from more or less arbitrary criteria, can be interpreted as probabilities in any meaningful sense. Moreover, given the enormous cost of developing and running a complex simulator, the number of such simulators available at any time will always be relatively small: treating a collection of weights as a probability distribution will therefore tend to underestimate uncertainty since alternative, as yet undeveloped, simulators could easily yield projections outside the currently available range.

(b) Probabilistic uncertainty assessment

In general, to quantify uncertainty probabilistically, it is necessary to use modern statistical techniques to derive probability density functions (PDFs) for quantities of interest. This approach requires the specification of a statistical model to represent the structure of the ensemble and the data it generates [13,15–22]. All of these references use a Bayesian framework and, in some instances, the statistical model implies a set of weights to be attached to the individual simulators: indeed, the work of Tebaldi *et al.* [15] was originally motivated by a desire to find the statistical model associated with the REA weights. However, difficulties remain with the interpretation of probabilities even from these sophisticated approaches. Fundamentally, this is because no systematic attempt has yet been made to sample representatively the range of possible modelling decisions that could be taken when constructing an MME. PDFs calibrated against existing MMEs are therefore not guaranteed to represent the full range of uncertainty, which has led to questions about the relevance of the resulting probabilities to the downstream user [23]. For an excellent discussion of other issues involved when interpreting probabilities in a closely related context, see Rougier [24].

The non-representative sampling of modelling decisions is partly due to the exchange of information between climate modelling groups. As a result, simulators share discrepancies with the actual climate [12,25]. Another source of shared discrepancies is that simulators are at best approximations of the real climate system: there is no guarantee that any collection of such approximations should be centred on reality. This has an important implication: if the set of potential climate simulators is centred on reality then projection uncertainty can in principle

be reduced to zero merely by including a large enough number of simulators, but otherwise uncertainty cannot be reduced indefinitely. Knutti *et al.* [25] provide evidence that this effect can be observed even with the relatively small number of climate simulators currently available. However, many of the existing statistical frameworks for the analysis of MMEs fail to account for shared discrepancies.

Alternative frameworks are available that, instead of assuming that an MME is centred on reality, rely on a concept of exchangeability: here, reality is treated essentially as though it were another simulator. Under this paradigm, increasing the number of simulators cannot reduce uncertainty indefinitely; however, the limiting factor is not the collective discrepancy but rather the variability of the distribution from which all of the simulators (and reality) are assumed to be drawn [26,27]. It seems, therefore, that although the existence of shared discrepancies in MMEs is generally recognized, it is not handled satisfactorily within current frameworks.

This paper aims to address all of the issues outlined above. The conceptual framework extends the work of Leith & Chandler [13] and is introduced in §2. A Bayesian approach is taken: given climate observations and the output from an MME, knowledge of the actual climate system is encapsulated in a posterior distribution for quantities of interest. Subsequently, §3 derives an explicit expression for this posterior distribution, in the specific situation where all quantities are Gaussian. Although the Gaussian assumption is slightly restrictive, it is still applicable in many situations where MMEs are used: moreover, the derived expressions provide insights that are expected to hold much more widely. In particular, some subjective judgements are inevitable in *any* attempt to interpret future projections: the use of a formal framework helps to clarify precisely what these judgements are, and hence to provide increased transparency in the ensuing analysis.

A fully Bayesian analysis would require the use of computationally intensive Markov chain Monte Carlo (MCMC) methods: this is not attempted here, however, in order to focus on the key concepts. Nonetheless, the mathematical analysis permits the development of a 'poor man's' implementation that circumvents the need for costly computational procedures, albeit by neglecting some estimation uncertainties: this is presented in §4, along with an application to the artificial data from figure 1. The use of an artificial example is quite deliberate: the aim is to focus attention on the main messages and concepts, whereas the use of a real example may detract from these by introducing application-specific issues. Applications involving simulations from the World Climate Research Program's Coupled Model Intercomparison Projects (CMIP3 and CMIP5) will be reported elsewhere.

2. Conceptual framework

Leith & Chandler [13] started by observing that climate projections aim to reproduce the statistical properties of the climate over extended time periods, rather than to reproduce the system trajectory in detail. Although such properties are often regarded just as convenient summaries of the system, more generally they can be considered as parameter estimates for a statistical model describing its outputs. Moreover, since in broad terms all simulators represent essentially the same dynamical processes, their outputs should have a similar statistical structure and hence can be described using the same form of statistical model. In this context, differences between the simulators (and between the simulators and the actual climate) correspond to different statistical model parameters.

To illustrate this concept, consider the artificial temperature series in figure 1. Each series shows a strong seasonal cycle oscillating about a fairly constant mean level; in the context of climate change, one might also be interested in the possible existence of linear time trends. Denote by Y_t the temperature at time t. Then the series structure could be summarized by fitting a regression model, of the form

$$Y_t = \beta_0 + \beta_1 \cos\frac{2\pi t}{365} + \beta_2 \sin\frac{2\pi t}{365} + \beta_3 t + \varepsilon_t, \tag{2.1}$$

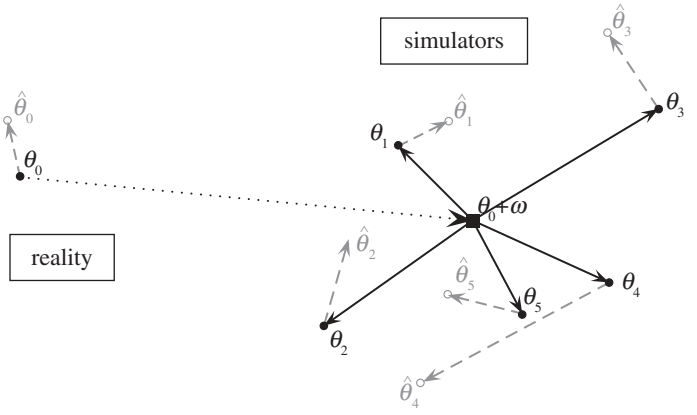

Figure 2. Schematic of the geometry for the proposed MME framework. $\boldsymbol{\theta}_0$ denotes the descriptor for the actual climate; $\{\boldsymbol{\theta}_i : i > 0\}$ are descriptors for simulators; and $\{\hat{\boldsymbol{\theta}}_i : i \geq 0\}$ are descriptor estimates obtained from data and simulator outputs. Grey dashed lines represent estimation errors; black dotted line represents shared simulator discrepancy, with simulator output descriptors centred on $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$. Arrows indicate direction of causal relationships in which an intervention at the 'parent' node is expected to produce a change at the 'child' node.

to each series separately, where the errors (ε_t) form a sequence of independent Gaussian random variables with mean zero and variance σ^2. Such a model, which aims to describe the structure of simulator outputs and the corresponding climate quantities, will be called a 'mimic' from here onwards.[1] In this example, the parameter β_0 controls the mean level of the series at time zero; β_1 and β_2 the amplitude and phase of the seasonal cycle; β_3 the time trend; and σ^2 the magnitude of any remaining irregular fluctuations. The parameters can be collected into a vector, say $\boldsymbol{\theta} = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \sigma^2)'$, which will be called a 'descriptor'. The differences between the three series suggest, however, that they cannot all be described using the same value of $\boldsymbol{\theta}$: rather, we should summarize the 'real climate' using a descriptor $\boldsymbol{\theta}_0$, and the simulator outputs using descriptors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively.

With this concept in mind, Leith & Chandler [13] suggested that uncertainty in an MME could be characterized by considering the descriptors $\{\boldsymbol{\theta}_i\}$ themselves to be drawn from some underlying joint probability distribution (thus acknowledging the potential for alternative simulators to yield descriptors outside the currently available range), and then by using the simulator outputs to learn about this underlying distribution. The resulting framework corresponds to a hierarchical model in which the outputs of the ith simulator (collected into a vector $\mathbf{Y}_i$, say) are considered to be generated from some stochastic process with parameter $\boldsymbol{\theta}_i$, and in which the $\{\boldsymbol{\theta}_i\}$ are themselves considered to be random.

Leith & Chandler [13] did not attempt to incorporate observations of actual climate into their analysis: as such, their work can be regarded as characterizing a notional population of simulators from which the MME is considered to be drawn. However, such a characterization provides little information about actual climate without making additional assumptions; for example, that the notional simulator population is centred on the actual climate in the sense that for $i \neq 0$ the expected value of $\boldsymbol{\theta}_i$ is $\boldsymbol{\theta}_0$. This approach was adopted by Furrer *et al.* [17] using essentially the same hierarchical model set-up as that described earlier. In practice, however, such an assumption is unrealistic in the presence of shared simulator discrepancies.

To make useful statements about the real climate system within this framework, two extensions are required. The first is to incorporate the actual climate descriptor $\boldsymbol{\theta}_0$ explicitly,

[1]The term 'mimic' is used by analogy with its meaning in ecology, where it denotes a species that has the appearance of another. In the present context, a model such as (2.1) is used to produce outputs that have the appearance of simulator outputs or of reality. Such models should be distinguished from statistical *emulators*, which aim to approximate the internal workings of a complex computer code rather than merely to mimic its outputs.

and the second is to acknowledge the potential for shared discrepancies among the simulators. Figure 2 illustrates how this might be achieved (see also [19,21]). This figure can be read as a geometrical representation of how the simulator descriptors relate to each other and to the real climate. Note in particular that the simulator descriptors are not centred upon $\boldsymbol{\theta}_0$ but rather upon the point $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$. The quantity $\boldsymbol{\omega}$, represented by the dotted line, is an ensemble-specific discrepancy that is shared by all simulators that could in principle have been included in the ensemble.

The grey components of figure 2 represent estimation errors arising from internal variability in the real and simulated climate systems: the quantities $\{\hat{\boldsymbol{\theta}}_i\}$ computed from data should be regarded as estimates of the underlying descriptors $\{\boldsymbol{\theta}_i\}$. From here onwards, we will consider that they are maximum-likelihood estimates (MLEs). Strictly speaking, we should write $\{\hat{\boldsymbol{\theta}}_i(\mathbf{Y}_i)\}$ to emphasize dependence on the data $\{\mathbf{Y}_i\}$. However, theoretical arguments [28, p. 246] suggest that in large samples the MLEs contain essentially all of the relevant information in the data regarding the $\{\boldsymbol{\theta}_i\}$: hence minimal information is lost by replacing each simulator's output with the corresponding MLE. This simplifies the discussion below, although the fundamentals would not change if using the raw data $\{\mathbf{Y}_i\}$ instead. In the current context, Leith & Chandler [13] suggested this approach as a means of reducing the computational burden of fitting a hierarchical model; they also provided empirical evidence that the information loss was indeed small in their application.

(a) The posterior distribution for $\boldsymbol{\theta}_0$

We can now assert that when interpreting the outputs from an MME, the aim is to use all of the available data to learn about the real climate descriptor $\boldsymbol{\theta}_0$. Having done this, if necessary the mimic can be used to generate pseudo-data with properties that are indistinguishable from those of the real climate system (at least, to the extent that the mimic represents the system structure). This can be seen as a sophisticated form of simulator bias correction: rather than use the simulator outputs directly, they are used indirectly to help calibrate a mimic of the real climate, and this mimic is used subsequently to inform, for example, strategies for mitigating and adapting to climate change.

Suppose now that there are m simulators in the ensemble so that, after reducing to MLEs, the data available consist of the descriptor estimates $\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m$. To establish how these estimates should be used to learn about $\boldsymbol{\theta}_0$, initially it is convenient to take a Bayesian approach whereby the required information is summarized in a posterior distribution: $\pi(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m)$ say, where $\pi(\cdot)$ denotes a generic probability density and the vertical bar '|' denotes a conditional distribution (in this case, conditioned on the values of $\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m$).

Here, the arrows in figure 2 become relevant. These represent causal linkages and point in the direction of causation so that an intervention in any node of the system is expected to lead to changes in its 'descendants'. Thus, for example, climate simulators are tuned to try and reproduce selected properties of the actual (historical) climate. One might reasonably judge that if the properties of the Earth's climate were different then the properties of the collective simulator outputs would also be different. This is represented by the arrow from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$ in figure 2. The lack of an arrow in the opposite direction indicates that changes in the simulators would not affect the real climate.

The reason for emphasizing these causal relationships is that they enable figure 2 to be interpreted as a directed acyclic graph from which probabilistic conditional independence relationships can be inferred (see the electronic supplementary material), and used to show that the posterior $\pi(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m)$ can be written as

$$\pi(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m) \propto \pi(\boldsymbol{\theta}_0)\pi(\hat{\boldsymbol{\theta}}_0|\boldsymbol{\theta}_0)\pi(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m|\boldsymbol{\theta}_0). \tag{2.2}$$

The first term in this expression represents the prior distribution for $\boldsymbol{\theta}_0$. The second is the joint density for the observed climate descriptor $\hat{\boldsymbol{\theta}}_0$ which, regarded as a function of $\boldsymbol{\theta}_0$, is the likelihood for $\boldsymbol{\theta}_0$ based solely on observations of the actual climate. Similarly, the final term is a likelihood

contribution from the simulator outputs. Thus, equation (2.2) can be paraphrased as

$$\text{Posterior} = \text{Prior} \times \text{Likelihood for observations} \times \text{Likelihood for simulator outputs.}$$

This looks slightly different from the usual 'Posterior = Prior × Likelihood' result found in most introductory treatments of Bayesian inference (e.g. [29, ch. 11]). However, the only difference is that in (2.2) the 'Likelihood' contribution is explicitly factorized into separate components corresponding to the real climate and to the simulators. Thus, although (2.2) has been derived from a Bayesian viewpoint, it can also be used as the basis for likelihood-based inference: if the prior $\pi(\boldsymbol{\theta}_0)$ is omitted then the remaining expression can be regarded as a likelihood function for $\boldsymbol{\theta}_0$ given the available data.

In a Bayesian framework, physical arguments can often be used to help choose the prior $\pi(\boldsymbol{\theta}_0)$ in (2.2) [13]. The prior also provides an opportunity to incorporate understanding of the climate system from sources other than the observations and simulators. For the observation component $\pi(\hat{\boldsymbol{\theta}}_0|\boldsymbol{\theta}_0)$, it will often be appropriate to exploit the fact that for many mimic structures, the distribution of the MLE is approximately multivariate normal (MVN) in large samples: $\hat{\boldsymbol{\theta}}_0 \sim \text{MVN}(\boldsymbol{\theta}_0, \mathbf{J}_0)$ say, where standard techniques can be used to estimate the covariance matrix $\mathbf{J}_0$ [29, §4.4]. The remaining factor $\pi(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m|\boldsymbol{\theta}_0)$ is most easily deduced from specifications of conditional distributions corresponding to the arrows in figure 2: each such conditional distribution represents the dependence of the 'offspring' node upon its parent, and it is here that the quantity $\boldsymbol{\omega}$ enters the calculations. Section 3 illustrates the idea.

At this point, it is appropriate to comment further on the role of $\boldsymbol{\omega}$. Note first that, although this was introduced to represent discrepancies that are shared between simulators, it also provides a means of handling non-representative sampling (see §1). At any given moment in time, one might envisage a notional population of climate simulators, defined by considering all of the possible modelling decisions that could be taken when developing such a simulator and that are consistent with current knowledge and computing power. Within this population, there will be collective discrepancies with the real climate system owing to the fact that simulators are at best approximations of reality. However, interaction between modelling groups, and the common practice of selecting specific simulators on the basis of their perceived suitability for particular applications, is such that any particular MME is unlikely to be representative of the population: the effect will be to introduce a systematic bias with respect to the population. The net effect remains, however, that the MME outputs are centred on a quantity other than $\boldsymbol{\theta}_0$: in the framework presented here, this quantity is denoted by $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$ so that $\boldsymbol{\omega}$ accounts simultaneously for both phenomena. There is, of course, a catch: since the non-representativeness now contributes to the shared discrepancy $\boldsymbol{\omega}$, it becomes necessary to make appropriate judgements about this quantity in order to make decision-relevant probability statements about reality. Although this may not be a trivial task, we will see below that the considerations required are intuitive and transparent. This at least provides the basis for an informed interpretation of probabilities computed from an MME.

Following from this, it is clear that $\boldsymbol{\omega}$ is ensemble-specific: as our knowledge and computing power evolve, so will the population of possible simulators. Moreover, the 'sampling bias' component of $\boldsymbol{\omega}$ depends on how ensembles are sampled from this population: this can be changed either by altering the way in which climate modelling groups operate, or by changing the criteria for including existing simulators in a particular ensemble. We return to this point later.

The idea of sampling from a notional population of simulators is conceptually convenient: it is perhaps most easily accepted when the simulators under consideration all belong to the same class; for example, when they are all coupled atmosphere–ocean general circulation models (AOGCMs) or when they are all energy balance models. An alternative interpretation of figure 2 centres around judgements of exchangeability of the simulators, conditional upon their consensus. In this case, the inclusion of $\boldsymbol{\omega}$ can be seen as a way of acknowledging that although the simulators may be exchangeable with each other, they probably are not exchangeable with the real climate system. For more discussion of exchangeability in the context of MMEs, see [30].

3. Mathematical analysis of the Gaussian case

To proceed beyond the general expression (2.2) for the posterior distribution of $\boldsymbol{\theta}_0$, it is necessary to make some assumptions about the distributions of the various quantities represented schematically in figure 2. This section examines the case when all of the required distributions are multivariate normal.

(a) The Gaussian specification

We have already seen that a multivariate normal assumption for the MLE $\hat{\boldsymbol{\theta}}_0$ will often be reasonable; the same argument can be used for the MLEs obtained from each of the simulators. Thus, we specify

$$\hat{\boldsymbol{\theta}}_i \sim \mathrm{MVN}(\boldsymbol{\theta}_i, \mathbf{J}_i) \quad (i = 0, \ldots, m), \tag{3.1}$$

so that $\mathbf{J}_i$ is the covariance matrix of the MLE for the ith data source. As discussed previously, this can be regarded as representing uncertainty attributable to internal variability. In cases where an individual simulator has been run several times for a single forcing scenario but with different initial conditions, the data from all runs can be pooled to yield a single MLE and associated covariance matrix (this assumes, however, that the system can be regarded as ergodic so that the properties of interest are unaffected by the choice of initial conditions: if this is not the case, it may be necessary to consider the choice of initial conditions as an additional level in the hierarchical framework).

Next, we specify a multivariate normal distribution for the simulator descriptors themselves. It is convenient to express this via discrepancies $\{\boldsymbol{\delta}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0 : i = 1, \ldots, m\}$. Specifically, we set

$$\boldsymbol{\delta}_i \sim \mathrm{MVN}(\boldsymbol{\omega}, \mathbf{C}_i) \quad (i = 1, \ldots, m), \tag{3.2}$$

so that the descriptors $\{\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \boldsymbol{\delta}_i\}$ are centred on a 'simulator consensus' $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$ as illustrated in figure 2. The multivariate normal assumption here may be considered unrealistic: heavy-tailed multivariate t distributions might be preferred, for example, to accommodate the possibility of outlying simulators. However, the use of simulator-specific covariance matrices $\{\mathbf{C}_i\}$ provides some flexibility to accommodate outlying simulators via distributions that are highly dispersed rather than heavy-tailed. The matrix $\mathbf{C}_i$ can be regarded as measuring the propensity for the ith simulator to deviate from the simulator consensus.

For the shared discrepancy $\boldsymbol{\omega}$, we set

$$\boldsymbol{\omega} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Lambda}). \tag{3.3}$$

It seems uncontroversial to assert here that $\boldsymbol{\omega}$ has expectation zero: without seeing any simulator outputs, it would be difficult to argue that they would be biased collectively in any particular direction with respect to the real climate. The multivariate normal assumption, however, is unverifiable: any MME yields only a single realized value of $\boldsymbol{\omega}$ so that it is impossible to verify the appropriateness of any distributional assumptions about it. In practice therefore, (3.3) is merely a convenient device that provides the flexibility to incorporate the shared discrepancy formally into an analysis.

The only remaining distribution to specify is the prior on $\boldsymbol{\theta}_0$. We set

$$\boldsymbol{\theta}_0 \sim \mathrm{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{3.4}$$

The specification is completed by stipulating that conditional on $\boldsymbol{\omega}$, the discrepancies $\{\boldsymbol{\delta}_i\}$ are mutually independent. This could be unrealistic, for example, if the same modelling group contributes more than one simulator to an MME or if some of the simulators are structurally identical but with different parametrizations (this occurs, for example, in perturbed physics ensembles). In principle, however, the potential for some of the $\{\boldsymbol{\delta}_i\}$ to be interdependent is

Table 1. Summary of covariance matrix notation.

notation	interpretation
$\mathbf{J}_i$	covariance matrix of $\hat{\boldsymbol{\theta}}_i$, representing uncertainty due to internal variability within data source i
$\mathbf{C}_i$	conditional (upon $\boldsymbol{\omega}$) covariance matrix of discrepancy $\boldsymbol{\delta}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0$, representing propensity for simulator i to deviate from the simulator consensus $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$
$\mathbf{D}_i$	defined as $\mathbf{C}_i + \mathbf{J}_i$
$\boldsymbol{\Lambda}$	covariance matrix of $\boldsymbol{\omega}$, representing propensity of simulators collectively to deviate from actual climate
$\boldsymbol{\Sigma}_0$	covariance matrix of prior distribution for $\boldsymbol{\theta}_0$

easily handled by extending the general framework. For example, if some modelling groups contribute several simulators to an MME then the different groups could be considered to define *families* of simulators, with each family being centred around its own consensus and the family consensuses themselves centred around $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$. Similarly, if some simulators have several *variants* (corresponding, for example, to different parametrizations), then each variant of the ith simulator will have its own descriptor and these variant descriptors will be centred on the simulator descriptor $\boldsymbol{\theta}_i$. Such structures can be represented by adding additional levels to the hierarchical framework: variant descriptors cluster around simulator descriptors, which cluster around family descriptors, which in turn are clustered around an overall simulator consensus. Clearly, however, this would substantially complicate the mathematical analysis, and we restrict attention here to the relatively simple structure illustrated in figure 2.

The covariance matrices $\{\mathbf{J}_i\}$, $\{\mathbf{C}_i\}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_0$ play a key role in the subsequent development, as do their inverses. For ease of reference therefore, their roles are summarized in table 1. Finally, it will be convenient to write $\mathbf{D}_i = \mathbf{C}_i + \mathbf{J}_i$ for $i = 1, \ldots, m$. Following standard statistical practice, the inverse of any covariance matrix will be referred to as a *precision matrix*.

(b) Limitations of the Gaussian specification

Although the Gaussian specification is widely applicable, it should not be used indiscriminately. Some potential limitations are indicated above. Another is that some descriptors by their nature cannot have Gaussian distributions: for example, variances must be non-negative, whereas Gaussian distributions extend over the entire real line. Such situations are most easily handled by working with one-to-one transformations of the descriptors: for example, rather than working with $\boldsymbol{\theta} = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \sigma^2)'$ in (2.1), one could work with $\boldsymbol{\theta}^* = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \log\sigma^2)'$.

There is a more subtle limitation to the Gaussian specification when studying several variables simultaneously. Here, any realistic mimic must preserve dependencies between the variables, often by specifying an inter-variable correlation matrix. In bivariate settings, transformations can be used to ensure that correlations lie in the range $(-1, 1)$: if ρ represents a generic correlation, one could work instead with a Gaussian specification for $\operatorname{arctanh}(\rho)$. However, for n variables there are $n(n-1)/2$ separate correlations and, if $n > 2$, these correlations must be mutually compatible in the sense that the correlation matrix is non-negative definite [29, p. 68]. This will generally not be the case if (transformed) correlations are sampled from a multivariate Gaussian distribution. An alternative would be to parametrize the correlation matrix and to treat the parameters rather than the correlations as descriptors: this would be natural when dealing with spatial or spatio-temporal data, for example. However, in some settings, it may be preferable to abandon the Gaussian specification altogether and to work directly with distributions, such as the Wishart and inverse Wishart families, that are designed for the modelling of covariance and correlation matrices (see [13] for a trivariate example).

(c) The posterior and its implications

With the Gaussian specification given by (3.1)–(3.4), it may be shown that the posterior distribution (2.2) for $\boldsymbol{\theta}_0$ is itself multivariate normal:

$$\left.\begin{aligned} &\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_m \sim \mathrm{MVN}(\boldsymbol{\tau},\mathbf{S}), \quad \text{where} \\ &\mathbf{S}^{-1}=\boldsymbol{\Sigma}_0^{-1}+\mathbf{J}_0^{-1}+\left[\boldsymbol{\Lambda}+\left(\sum_{k=1}^{m}\mathbf{D}_k^{-1}\right)^{-1}\right]^{-1} \\ \text{and} \quad &\boldsymbol{\tau}=\mathbf{S}\left[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\mathbf{J}_0^{-1}\hat{\boldsymbol{\theta}}_0+\left(\mathbf{I}+\sum_{k=1}^{m}\mathbf{D}_k^{-1}\boldsymbol{\Lambda}\right)^{-1}\sum_{i=1}^{m}\mathbf{D}_i^{-1}\hat{\boldsymbol{\theta}}_i\right]. \end{aligned}\right\} \tag{3.5}$$

Here, $\mathbf{I}$ denotes the $p \times p$ identity matrix, where p is the number of elements in $\boldsymbol{\theta}_0$. The derivation of this result is given in the electronic supplementary material.

From a Bayesian viewpoint, the mean of a posterior distribution is the optimal estimator of the unknown parameter vector in the sense of minimizing the expected squared estimation error [31, §3.2]. Inspection of (3.5) reveals that in the present setting the posterior mean $\boldsymbol{\tau}$ is a matrix-weighted average of the prior mean $\boldsymbol{\mu}_0$, the estimated descriptor $\hat{\boldsymbol{\theta}}_0$ obtained from observations of actual climate and the estimated descriptors $\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_m$ from each of the simulators. A consequence of the matrix-valued weights is that the relative weights attached to the different information sources will typically vary between components of $\boldsymbol{\theta}_0$. This ensures, for example, that each simulator contributes to the posterior mean for components where it is informative but not elsewhere: simulators' strengths are exploited while their weaknesses are discounted. This immediately resolves one of the difficulties associated with simulator weighting, discussed in §1.

Recall from table 1 that $\mathbf{D}_i = \mathbf{C}_i + \mathbf{J}_i$. Thus, the weighting matrices in the posterior mean $\boldsymbol{\tau}$ depend on measures of internal variability (the $\{\mathbf{J}_i\}$), simulator consensus (the $\{\mathbf{C}_i\}$) and shared discrepancy ($\boldsymbol{\Lambda}$). This seems intuitively reasonable. It may seem surprising, however, that the weighting matrices do not depend explicitly on the observed performance of the different simulators: rather, the covariance matrices all relate to *expected* variation. Thus, for example, if the $\{\mathbf{D}_i\}$ are all equal, the simulators all receive the same weighting matrix regardless of how well they reproduce features of historical climate. This seems counterintuitive, but figure 2 provides some insight into the result. Here, simulator 2 yields the descriptor estimate closest to that for the real climate, but this is coincidental: the simulators collectively provide information about their consensus $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$ rather than directly about $\boldsymbol{\theta}_0$ and, given that they are all different, it is guaranteed that one of them will be closer to reality than the others.

From this, it may be tempting to conclude that historical simulator performance plays no role in the interpretation of future projections. This would be incorrect, however. In practice, historical performance is the basis for estimating the matrices $\{\mathbf{C}_i\}$ and $\boldsymbol{\Lambda}$ (see §4). Historical performance may also be used to inform the choice of simulators to include in an ensemble: following the earlier discussion of sampling bias, the selection of simulators on the basis of historical performance can be regarded as an attempt to reduce the magnitude of the shared discrepancy $\boldsymbol{\omega}$, which will be reflected in the value of $\boldsymbol{\Lambda}$.

Result (3.5) also sheds some light on the Bayesian framework of Tebaldi *et al.* [15], who proposed in the context of scalar quantities that each simulator's output should be drawn from a distribution with a different variance. According to (3.5), this is one of only two ways to ensure the differential weighting of simulators when calculating a posterior mean (which was a key objective for Tebaldi *et al.* in trying to replicate the REA method, as discussed in §1): the other is via the representation of simulator internal variability, regarding which there is little scope for flexibility since the matrices $\{\mathbf{J}_i\}$ are determined by the standard theory of maximum likelihood.

We turn next to the expression for the precision matrix $\mathbf{S}^{-1}$ in (3.5). This expression provides insights into the contributions of different types of information to knowledge of actual climate, since a 'large' value of this precision matrix corresponds to high certainty about $\boldsymbol{\theta}_0$. Note first that

if we could increase the number of relevant observations of actual climate indefinitely, $\mathbf{J}_0$ would tend to a zero matrix since these observations would effectively provide perfect information about $\boldsymbol{\theta}_0$. In this case, regardless of the simulator outputs and prior specification, the precision matrix $\mathbf{S}^{-1}$ in (3.5) would become arbitrarily large due to the contribution from $\mathbf{J}_0^{-1}$, and our uncertainty could be reduced indefinitely. To paraphrase: if we had perfect knowledge of reality, climate simulators would be redundant. This provides a helpful sanity check on the result.

Consider next the implications of increasing the number of simulators, with $\mathbf{J}_0$ and $\boldsymbol{\Sigma}_0$ held fixed. The simulators collectively contribute $[\boldsymbol{\Lambda} + (\sum_{k=1}^{m} \mathbf{D}_k^{-1})^{-1}]^{-1}$ to the precision matrix in (3.5). Moreover, the $\{\mathbf{D}_k\}$ (and hence their inverses) are all positive definite since they are sums of covariance matrices (table 1). Thus, as $m \to \infty$, the simulator contribution cannot exceed $\boldsymbol{\Lambda}^{-1}$ and the maximum attainable precision is $\boldsymbol{\Sigma}_0^{-1} + \mathbf{J}_0^{-1} + \boldsymbol{\Lambda}^{-1}$. Effectively, the precision is limited by the shared discrepancy $\boldsymbol{\omega}$.

Similar arguments can be used to deduce the effects of increasing the length and number of runs of individual simulators (which corresponds to letting the $\{\mathbf{J}_i\}$ tend to zero), and of increasing simulator consensus (by letting the $\{\mathbf{C}_i\}$ tend to zero): in neither case is it possible to increase the precision beyond $\boldsymbol{\Sigma}_0^{-1} + \mathbf{J}_0^{-1} + \boldsymbol{\Lambda}^{-1}$. To paraphrase again: in the presence of discrepancies that are shared between simulators, uncertainty cannot be reduced indefinitely by increasing the number of simulators, the length and number of runs or by working to achieve greater inter-simulator consensus. Other authors have noted this previously, on both theoretical and empirical grounds (see §1). However, as far as we are aware this is the first time that the relevant issues have been quantified precisely within such a general mathematical framework.

These results also have implications for the design of future MMEs. By quantifying the contributions of the various sources of uncertainty to $\mathbf{S}^{-1}$, it is possible in principle to compare the potential information gains from different design strategies: for example, to determine whether it is better to carry out many runs of a few simulators, or fewer runs of more simulators. Moreover, the results show that in the absence of unlimited observations of the real climate, the shared discrepancy $\boldsymbol{\omega}$ is the key barrier to reducing uncertainty. This is discussed further in §5.

The perspective so far has been Bayesian. However, as noted earlier, the likelihood function for $\boldsymbol{\theta}_0$ is obtained by omitting the prior $\pi(\boldsymbol{\theta}_0)$ from the posterior density. In the context of equation (3.5), this can be achieved equivalently by setting the prior precision $\boldsymbol{\Sigma}_0^{-1}$ to a zero matrix. In this case, the MLE of $\boldsymbol{\theta}_0$ based on all of the available data is the mode of the likelihood function, which is $\boldsymbol{\tau}$; moreover, the covariance matrix of the MLE is $\mathbf{S}$ by the usual duality with Bayesian posterior covariance matrices [29, §11.2]. Thus, a likelihood-based analysis leads to exactly the same insights as the Bayesian approach above.

(d) Assessing changes in climatological summaries

Many studies are concerned with projected changes in climatological summaries such as regional mean temperatures or extreme precipitation frequencies. This kind of problem can be dealt with in the proposed framework by considering the descriptor vector $\boldsymbol{\theta} = (\boldsymbol{\psi}'_{\text{historical}}\ \boldsymbol{\psi}'_{\text{future}})'$, where $\boldsymbol{\psi}$ is a vector containing the climatological summaries of interest. In principle, the theory can be applied straightforwardly in this situation, although the absence of observations on the real climate of the future leads to some challenging problems of estimation in practice (see §4).

It is also worth considering how to define the matrix $\mathbf{J}_0$ in this situation, because the components of $\hat{\boldsymbol{\theta}}_0$ corresponding to $\boldsymbol{\psi}'_{\text{future}}$ are undefined. Note, however, that these components can equivalently be considered as estimated with zero precision. The precision matrix for $\hat{\boldsymbol{\theta}}_0$ can therefore be partitioned into blocks corresponding to historical and future components, as

$$\mathbf{J}_0^{-1} = \begin{pmatrix} \mathbf{J}_{\text{historical}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\text{future}}^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{J}_{\text{historical}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{3.6}$$

in an obvious notation.

4. Poor man's version

In practical applications, the values of the various covariance matrices in table 1 will be unknown; they must therefore be estimated in order to apply the results of §3. The 'gold standard' for addressing this problem is a fully Bayesian treatment that specifies prior distributions for these covariance matrices and modifies the posterior (3.5) appropriately to accommodate the additional uncertainty due to the estimation. In this case, however, in general an exact analytical treatment seems infeasible so that the posterior must be computed using alternative methods, for example, using MCMC techniques [29, §11.3]. Such techniques are often computationally intensive; moreover, their correct use requires good awareness of the underlying theory and potential pitfalls. This section therefore presents a 'poor man's version' that is easy and cheap to implement and may be suitable for routine implementation in the climate impacts community. The approach can be regarded as a form of empirical Bayes analysis (e.g. [29, §11.5]).

(a) Cutting corners

The proposal here is to estimate all of the required covariance matrices directly from the available data, and to plug these estimates into (3.5) without accounting for the estimation uncertainty. In practice, this uncertainty could be large, especially for the matrix $\boldsymbol{\Lambda}$ (see below). Moreover, the proposal relies critically upon the Gaussian framework developed in §3, with the attendant limitations discussed there. Nonetheless, the approach is likely to improve upon much current practice in which uncertainty assessments for MMEs are based on heuristic weighting schemes with little formal justification.

In general, to ensure that the 'plug-in' estimate of $\mathbf{S}$ in (3.5) is strictly positive definite, and hence that the proposed procedure delivers a valid joint probability distribution for $\boldsymbol{\theta}_0$, it is necessary that the various covariance matrix estimates involved in the calculation of $\mathbf{S}$ are themselves non-negative definite. All of the proposals below meet this requirement.

As noted previously, the $\{\mathbf{J}_i\}$ can usually be obtained directly from statistical software output when the mimics are fitted. For the 'simulator consensus' covariance matrices $\{\mathbf{C}_i\}$, it will often be plausible to consider that the available simulators can be partitioned into one or more subsets such that all simulators within a subset are equally credible; this implies that they share a common covariance matrix. The shared covariance matrix for some subset, $\mathcal{S}$ say, can then be estimated from the estimated descriptors within that subset as

$$\hat{\mathbf{C}}_{\mathcal{S}} = \frac{1}{m_{\mathcal{S}} - 1} \sum_{i \in \mathcal{S}} (\hat{\boldsymbol{\theta}}_i - \hat{\bar{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_i - \hat{\bar{\boldsymbol{\theta}}})'. \tag{4.1}$$

Here, $m_{\mathcal{S}}$ is the number of simulators in $\mathcal{S}$ and $\hat{\bar{\boldsymbol{\theta}}} = m^{-1} \sum_{i=1}^{m} \hat{\boldsymbol{\theta}}_i$ is the overall mean of the estimated simulator descriptors.

To identify subsets of the simulators with a common covariance matrix, it is tempting to consider groups of simulators that tend to agree more or less well with the overall consensus. This is incorrect, however. To see why, consider an MME consisting of a mixture of oversimplified and coarse-resolution simulators, together with state-of-the-art AOGCMs. One might expect the coarse-resolution simulator descriptors to be scattered widely about the overall consensus, whereas the AOGCM descriptors would be more tightly clustered. However, one or two coarse-resolution simulators could produce outputs that are close to the consensus by chance: it would not be appropriate, however, to consider these as equivalent in some sense to the AOGCMs. Rather, the subsets must be chosen on the basis of expert judgements about the simulators themselves. Such judgements might be based upon considerations of resolution and the representation of key dynamical processes—in the first instance for example, one might simply allocate different 'generations' of simulators to different subsets.

In fact, (4.1) will tend to overestimate $\mathbf{C}_{\mathcal{S}}$. There are two reasons for this. The first is the use of the descriptor estimates $\{\hat{\boldsymbol{\theta}}_i\}$ rather than the (unknown) true descriptors $\boldsymbol{\theta}_i$. The precise magnitude of this effect can be determined by considering the case when all of the available simulators fall in

the same subset, so that there is a single consensus matrix, $\mathbf{C}$ say. In the electronic supplementary material, it is shown that in this case the expected value of (4.1) is $m^{-1}\sum_{i=1}^{m}\mathbf{J}_i + \mathbf{C} = \bar{\mathbf{J}} + \mathbf{C}$, say. If $\bar{\mathbf{J}}$ is small by comparison with $\mathbf{C}$ therefore, this particular source of overestimation is unimportant. This seems to be the case in general: for the artificial example analysed below, the largest eigenvalues of (4.1) and $\bar{\mathbf{J}}$ are 24.12 and 0.02, respectively, suggesting that this source of overestimation is negligible here. Similar results have been obtained in real examples, to be reported elsewhere.

In situations where $\bar{\mathbf{J}}$ is not small by comparison with $\mathbf{C}$, the result in the previous paragraph suggests that an improved estimator of $\mathbf{C}$ is

$$\frac{1}{m-1}\left[\sum_{i=1}(\hat{\boldsymbol{\theta}}_i - \hat{\bar{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_i - \hat{\bar{\boldsymbol{\theta}}}) - \bar{\mathbf{J}}\right].$$

However, this estimator has the disadvantage that unlike (4.1), in general it fails to deliver a non-negative definite result so that the corresponding estimate of $\mathbf{S}$ is not a valid covariance matrix. In such situations, one must be prepared either to accept the bias inherent in the estimator (4.1), to consider increasing the length or number of simulator runs so as to reduce the magnitude of $\bar{\mathbf{J}}$ (this shows, incidentally, that the value of increasing the length or number of simulator runs should not be seen solely in terms of the contribution to the posterior (3.5)), or to carry out a fully Bayesian analysis using MCMC techniques as described at the start of this section.

The second source of overestimation in (4.1) is that if the simulators are partitioned into $S > 1$ subsets, the individual simulator descriptors within subset $\mathcal{S}$ will deviate more from the overall mean $\bar{\boldsymbol{\theta}}$ than from the subset-specific mean $m_{\mathcal{S}}^{-1}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{\theta}}_i$. Once again, however, these effects are likely to be small and, in any case, can be regarded as offsetting slightly the overall underestimation of uncertainty in the posterior due to the use of plug-in covariance matrix estimates.

Finally, we consider the estimation of $\boldsymbol{\Lambda} = \mathrm{E}(\boldsymbol{\omega}\boldsymbol{\omega}')$. If the historical observations provide information, however limited, about all components of $\boldsymbol{\theta}_0$, then an estimate of $\boldsymbol{\omega}$ can be obtained as $\hat{\boldsymbol{\omega}} = \hat{\bar{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}}_0$ and $\boldsymbol{\Lambda}$ can be estimated as its empirical counterpart: $\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\omega}}\hat{\boldsymbol{\omega}}'$. Clearly, this estimate will be very imprecise: one might reasonably expect that the associated uncertainty will be the main source of differences between the procedure developed here and a full Bayesian analysis (which would allow the incorporation of additional information about simulator performance and its relation to reality, by specifying an appropriate prior distribution for $\boldsymbol{\Lambda}$). Nonetheless, as we demonstrate below, the current procedure produces appealing and intuitive results.

If the descriptor vector contains components that cannot be estimated from historical observations, the corresponding components of the shared discrepancy $\boldsymbol{\omega}$ clearly cannot be estimated and hence the above suggestion for estimating $\boldsymbol{\Lambda}$ cannot be used. To indicate how one might proceed in such circumstances, consider once again the case $\boldsymbol{\theta} = (\boldsymbol{\psi}'_{\text{historical}}\ \boldsymbol{\psi}'_{\text{future}})'$ in which the descriptor vector contains 'historical' and 'future' values of climatological summaries. In this situation, the procedure outlined above could be used to estimate the block of $\boldsymbol{\Lambda}$ corresponding to the $\boldsymbol{\psi}_{\text{historical}}$-components of $\boldsymbol{\omega}$: $\hat{\boldsymbol{\Lambda}}_{\text{historic}} = \hat{\boldsymbol{\omega}}_{\text{historic}}\hat{\boldsymbol{\omega}}'_{\text{historic}}$, say. No data are available to estimate the remainder of the matrix, however: this must therefore be specified entirely on the basis of judgement. One possibility is to assert that the shared discrepancy will be the same in the future as it was in the past: $\boldsymbol{\omega}_{\text{future}} = \boldsymbol{\omega}_{\text{historic}}$. Perhaps more realistically given that simulators are typically tuned to reproduce some aspects of historical climate, one might anticipate that shared discrepancies could change in the future (see [19] for some discussion of this): $\boldsymbol{\omega}_{\text{future}} = \boldsymbol{\omega}_{\text{historic}} + \boldsymbol{\eta}$ say. One possible judgement about $\boldsymbol{\eta}$ is that it is a zero-mean vector, independent of $\boldsymbol{\omega}_{\text{historic}}$, with covariance matrix $K\boldsymbol{\Lambda}_{\text{historic}}$ for some constant $K > 0$. In this case, we would have

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{\text{historic}} & \boldsymbol{\Lambda}_{\text{historic}} \\ \boldsymbol{\Lambda}_{\text{historic}} & (1+K)\boldsymbol{\Lambda}_{\text{historic}} \end{pmatrix}, \tag{4.2}$$

which can be estimated by replacing $\boldsymbol{\Lambda}_{\text{historic}}$ with $\hat{\boldsymbol{\Lambda}}_{\text{historic}}$ throughout. Equation (4.2) represents a judgement that shared discrepancies are likely to increase in magnitude in the future (the

value of K controls the extent of this increase), but that their relative magnitudes for different components of $\boldsymbol{\psi}$ are likely to be similar to those in the past. Of course, as a judgement regarding the extent to which historical performance is relevant in assessing the future, the choice of K is inevitably subjective (see the discussion towards the end of §1): in any analysis therefore, it is worth exploring the sensitivity of results to the choice of K. This is illustrated in §4*b*.

(b) Example

For illustrative purposes, we now return to the artificial example from figure 1. The mimic (2.1) is fitted separately to the 'observations' and to the outputs from each simulator. For the simulator outputs, the mimic is fitted separately to data from the 'historical' period 1980–2010 for which observations are available, and to the 'future' period 2011–2050. This example therefore falls into the framework considered above, in which the descriptor vector contains historical and future values of climatological summaries. The time origin is taken as 1980 in both periods, to ensure that the historical and future descriptor components represent the same quantities. Moreover, time is measured in decades so that the coefficient β_3 measures the mean change per decade.

For regression models such as (2.1), the MLEs of the regression coefficients are in fact the usual least-squares estimates: the corresponding covariance matrices are easily estimated using standard least-squares theory [29, §8.3]. The MLE for the residual variance in a regression model is $n^{-1}\sum_{t=1}^{n} e_t^2$, where n is the sample size and e_t is the tth residual. However, this variance estimator is (slightly) biased and instead it is standard practice to use the unbiased estimator $(n-q)^{-1}\sum_{t=1}^{n} e_t^2$, where q is the number of regression coefficients estimated. This unbiased estimator is just a scaled version of the MLE: we have therefore used it here, since any one-to-one transformation of the MLE contains the same information as the MLE itself. The estimator is uncorrelated with the estimated regression coefficients; moreover, its distribution is proportional to chi-squared on $(n-q)$ degrees of freedom and, in large samples, can be approximated as normal with mean σ^2 and variance $2\sigma^4/(n-q)$. Thus, for $i=0,1,2$, the 'historical' block of $\mathbf{J}_i$ is a $(q+1)\times(q+1)$ matrix containing the estimated covariance matrix of the historical regression coefficients in the first q rows and columns, and with zeroes in the final row and column except for the $(q+1,q+1)$ position which contains the historical estimate of $2\sigma_i^4/(n-q)$. Similar estimates are used for the 'future' blocks of $\mathbf{J}_1$ and $\mathbf{J}_2$; historical and future components are taken as independent so that the remaining blocks of these matrices are zero. The lack of future observations means, however, that the future block of $\mathbf{J}_0$ cannot be estimated; its inverse is therefore set to zero as in equation (3.6).

The descriptor estimates for each data source are given in table 2*a*. Simulator 1 (respectively, 2) yields a much lower historical estimate of β_0 (respectively, β_1) than the observations. This reflects the tendency for simulator 1 to be too cool, and for simulator 2 to underestimate the amplitude of the seasonal cycle (recall equation (2.1)).

Table 2*b* gives the mean $\hat{\boldsymbol{\theta}}$ of the estimated simulator descriptors, along with the estimated discrepancy $\hat{\boldsymbol{\omega}}_{\text{historic}}$. The two simulators are regarded as equally credible so that $\mathbf{C}_1=\mathbf{C}_2=\mathbf{C}$, say: thus, $\mathbf{C}$ is estimated using (4.1), with $\mathcal{S}=\{1,2\}$ and $m_{\mathcal{S}}=2$. The estimated discrepancy $\hat{\boldsymbol{\omega}}_{\text{historic}}$ is used to construct an estimate of $\boldsymbol{\Lambda}$ as in the discussion leading to equation (4.2). For illustrative purposes, we consider three separate values of K in that equation: $K=0, 0.2$ and 1. Setting $K=0$ amounts to a judgement that shared simulator discrepancies will not change in the future; $K=0.2$ corresponds to the potential for the magnitude of these discrepancies to increase slightly; and $K=1$ could be considered as representing scepticism regarding the extent to which historical simulator runs are informative about future climate.

Finally, to compute the posterior (3.5), it is necessary to specify the prior mean and covariance matrix. For the sake of simplicity, we assert complete prior ignorance by specifying $\boldsymbol{\Sigma}_0^{-1}=\mathbf{0}$. In this case, the prior contributes nothing to either the posterior mean or variance. If preferred therefore, these may be interpreted as resulting from a likelihood-based rather than a Bayesian analysis as discussed in §3.

Table 2. Analysis of the artificial temperature data from figure 1. (*a*) Descriptor estimates in model (2.1) fitted to 'observations' ($\hat{\boldsymbol{\theta}}_0$) and to output from two simulators. (*b*) Estimates of simulator consensus $\hat{\bar{\boldsymbol{\theta}}}$ and shared discrepancy $\boldsymbol{\omega}$. (*c*) Posterior means and standard deviations, obtained from (3.5) with plug-in covariance matrix estimates and $\boldsymbol{\Lambda}$ estimated according to (4.2) with different values of K.

	historical					future				
	β_0	β_1	β_2	β_3	σ^2	β_0	β_1	β_2	β_3	σ^2
(*a*) estimates										
$\hat{\boldsymbol{\theta}}_0$	11.98	5.10	−0.03	0.09	1.00	—	—	—	—	—
$\hat{\boldsymbol{\theta}}_1$	7.99	4.73	−0.34	0.08	1.00	7.57	4.66	−0.39	0.19	1.01
$\hat{\boldsymbol{\theta}}_2$	11.56	1.65	0.11	0.02	0.87	11.52	1.49	0.10	0.02	0.89
(*b*) consensus										
$\hat{\bar{\boldsymbol{\theta}}}$	9.78	3.19	−0.12	0.05	0.94	9.55	3.07	−0.15	0.11	0.95
$\hat{\boldsymbol{\omega}}$	−2.20	−1.91	−0.09	−0.04	−0.06	—	—	—	—	—
(*c*) posterior										
$\boldsymbol{\tau}$	11.98	5.10	−0.03	0.09	1.00	11.75	4.98	−0.06	0.14	1.02
s.d. ($K = 0$)	0.08	0.07	0.04	0.03	0.04	0.16	0.10	0.05	0.03	0.04
s.d. ($K = 0.2$)	0.08	0.07	0.04	0.03	0.04	1.00	0.86	0.06	0.03	0.05
s.d. ($K = 1$)	0.08	0.07	0.04	0.03	0.04	2.21	1.91	0.10	0.05	0.08

Table 2*c* gives the posterior mean $\boldsymbol{\tau}$, as well as the posterior standard deviations (i.e. the square roots of the diagonal elements of $\mathbf{S}$ in (3.5)) for each of the considered values of K. The posterior means are in fact identical for all choices of K, so these are presented just once. The posterior mean for the historical components of the descriptor vector appears identical to the historical estimator $\hat{\boldsymbol{\theta}}_0$; in fact, however, there are small differences in the third decimal place (not shown). Of more interest are the results for the future components: here, the posterior mean for β_0 is 11.75 which is considerably greater than the simulator consensus 9.55. In fact, the difference between the two is 2.20 which, to two decimal places, agrees with the estimate of the shared discrepancy for the same quantity during the historical period (once again, there are small differences in the third decimal place). Similar relationships hold for the other components. Essentially, this is due to the structure of the $\boldsymbol{\Lambda}$ matrix which, for all values of K, asserts that future discrepancies are likely to be in the same direction as historical ones. Concentrating for the moment on β_0 and β_1, the net effect is that the posterior mean for β_0 is close to the value obtained from simulator 2, whereas that for β_1 is close to the value obtained from simulator 1. The analysis therefore appears to exploit the strengths of each simulator while simultaneously discounting their weaknesses: remarkably, however, this is achieved without any differential treatment of the simulators. The explanation for this is that the relative configurations of historical and future components of the simulator descriptors are similar, and the future consensus is judged to stand in similar relation to reality as the historical one.

The posterior standard deviations in table 2 are small for the historical components of the descriptor: this is because the historical observations are informative for these components. For the future components, however, the standard deviations are typically much larger and are strongly dependent on the choice of K in (4.2) (see the earlier discussion highlighting the need to assess sensitivity to this choice). This is essentially due to uncertainty about future changes in the shared simulator discrepancy. Note, however, that the standard deviations increase with K at different rates for different descriptor components: those for β_0 and β_1 increase more than 10-fold as K changes from 0 to 1, whereas those for the remaining components merely double. This presumably reflects the relative variability of the historical values of the various components:

overall, the inter-simulator agreement, as well as the agreement between simulators and reality, is greater for the remaining components than for β_0 and β_1.

In §1, this example was introduced in the context of a hypothetical problem involving the length of growing season for a particular crop. To use the results of this analysis in such a problem, one would first sample multiple descriptor vectors from the posterior distribution (3.5); then use each of these descriptor vectors to simulate a temperature sequence from the mimic (2.1); and finally, for each simulated sequence, calculate the annual time series of growing season lengths. For each year of simulation, the resulting collection of simulated growing season lengths could then be treated as a probability distribution, representing uncertainty due both to the internal variability in temperature sequences and to the shared discrepancy and lack of consensus between simulators.

5. Summary

The proposed framework handles most, if not all, of the issues discussed in §1 relating to the analysis of MMEs. The key innovations are the explicit representations of reality ($\boldsymbol{\theta}_0$) and of shared simulator discrepancies. Other frameworks can be regarded as special cases of that presented here: for example, the approach of Tebaldi *et al.* 2005 [15] and Smith *et al.* [18] is obtained by setting the covariance matrix $\boldsymbol{\Lambda}$ of the shared discrepancy to zero, and by defining individual 'consensus' matrices $\{\mathbf{C}_i\}$ for each simulator.

Although the framework introduced in §2 is completely generic in principle, arguably the most useful insights have been obtained from the closed-form solution (3.5) to the simplified version in §3. The simplification, which involves assumptions of Gaussianity and conditional independence of the simulators, is nonetheless sufficiently realistic that the insights might be expected to hold much more generally. In particular, the assumption of Gaussianity is not necessary: in non-Gaussian settings, $\boldsymbol{\tau}$ and $\mathbf{S}$ in (3.5) can be regarded as emanating from a Bayes linear analysis [32, ch. 3]: this shows that $\boldsymbol{\tau}$ provides in some sense the optimal linear combination of information from different sources, regardless of the exact distributional assumptions.

As discussed in §3, the assumption of conditional independence (i.e. that the $\{\boldsymbol{\delta}_i\}$ in (3.2) are mutually independent) will often be less realistic. However, this is unlikely to affect the broad conclusions presented here, and the framework can in principle be extended to accommodate the kinds of structures encountered in practice.

This work challenges several aspects of current thinking regarding the analysis of MMEs. In particular, it shows that simulator outputs should not be weighted based on metrics of historical performance. This conclusion is perhaps unexpected, although it seems consistent with the emerging empirical evidence that unweighted averages often perform favourably by comparison with metric-based weighting schemes. In view of this, it is worth emphasizing that any disagreement with the conclusions must ultimately centre on whether figure 2, or something similar to it, is a plausible representation of the structure of an MME.

In addition to the role of simulator weighting, two further insights have emerged. The first is the nature of the subjective judgements that are required, in particular regarding the future evolution of shared discrepancies between climate simulators and reality. The ability to specify such judgements in an intuitive manner, via the covariance matrix $\boldsymbol{\Lambda}$, is an appealing feature. This has been illustrated in the argument leading to (4.2), in which the quantity K can be regarded as a convenient means of representing the analyst's faith in the future performance of the simulators. In operational terms, K plays much the same role as the θ parameter in Tebaldi *et al.* [15] and Smith *et al.* [18], which the latter authors describe as representing 'the differential between the reliabilities of current and future model projections'. If the analyst is uncomfortable with the sensitivity of the results to the choice of K here, a fully Bayesian approach could proceed by assigning a hyperprior distribution so that the posterior reflects the uncertainty in this choice: indeed, this approach is taken by Tebaldi *et al.* [15] and Smith *et al.* [18] with respect to their θ parameters. Even in this case, however, it is likely that the results will be sensitive to the choice of hyperprior: for an illustration of this, see fig. 11 of Buser *et al.* [19].

The final insight relates to the design of future MMEs, since the posterior (3.5) provides a way to quantify the benefits of increasing the precision of various components of the ensemble and hence to decide how best to allocate the available computing resources (as noted in §4, it is necessary also to consider the precision with which the various covariance matrices can be estimated). Similar comments have been made by Goldstein & Rougier [33], in connection with ensembles obtained by varying the inputs to a single simulator. In particular, the discussion in §3 shows that the main barrier to reducing uncertainty about future climate is the shared discrepancy $\boldsymbol{\omega}$. If uncertainty is to be reduced therefore, future MMEs must be designed so as to bring $\boldsymbol{\omega}$ closer to $\mathbf{0}$. One way to achieve this is to use historical performance to inform the selection of simulators for use in any particular application. A more ambitious option is to develop new simulators that are deliberately dissimilar to those currently available so as to centre the notional 'simulator population' more closely upon reality.

In terms of implementation, perhaps the most difficult aspect of the suggested framework is the specification of a statistical mimic. In any application, it will be necessary to consider carefully what features of the climate system are critical and to develop a mimic that reproduces these well, potentially at the expense of less important features. Moreover, statistical mimics may struggle to represent feedbacks between components of the climate system. Nonetheless, in applications one is typically interested only in a small part of the entire system, and this often simplifies the task of finding mimics that reproduce the main structures of interest adequately for practical purposes. In settings where interest lies merely in time-averaged global or regional averages, for example, it will often be adequate to consider that these averages are drawn from underlying normal distributions so that the descriptor vectors consist solely of means and variances for these distributions—this is much simpler than the example considered in this paper. More generally, application-specific statistical models are often used in the climate impacts community, and statistical techniques can reproduce a wide variety of impact-relevant properties: see, for example, the review of precipitation downscaling techniques in Maraun *et al.* [34]. Maximum-likelihood estimation for many of these techniques is straightforward, as is sampling from the associated mimic: this provides a means of obtaining probabilistic projections for use in impacts studies, as discussed in §4. For mimics where likelihood-based descriptor estimation is infeasible, other estimators may be used instead: the only price to pay for this is that typically there will be some information loss in moving from the original data to the descriptor estimates.

The illustrative example considered in this paper was deliberately artificial and was kept simple in order to focus on fundamental concepts. Ongoing work seeks to investigate the performance of the approach in practical applications involving real MMEs and potentially more complex systems, where the implementation may be substantially more complicated; and to compare the 'poor man's' version with a fully Bayesian analysis. Fully Bayesian implementations are not trivial, and experience will be required to highlight potential difficulties and ways of overcoming them. The results of this work will be reported elsewhere.

This research was started while the author was visiting the Isaac Newton Institute for Mathematical Sciences in Cambridge in November 2010, as part of the programme Mathematical and Statistical Approaches to Climate Modelling and Prediction. The support of the Institute is gratefully acknowledged, as are stimulating discussions with several programme participants, in particular Jonty Rougier. Finally, the paper has benefited from the constructive input of several reviewers: their time and patience is appreciated.

References

1. Hawkins E, Sutton R. 2009 The potential to narrow uncertainty in regional climate predictions. *Bull. Am. Meteorol. Soc.* **90**, 1095–1107. (doi:10.1175/2009BAMS2607.1)
2. Tebaldi C, Knutti R. 2007 The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. R. Soc. A* **365**, 2053–2075. (doi:10.1098/rsta.2007.2076)
3. Knutti R. 2010 The end of model democracy? *Climatic Change* **102**, 395–404. (doi:10.1007/s10584-010-9800-2)

4. Giorgi F, Mearns LO. 2002 Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the 'Reliability Ensemble Averaging' (REA) method. *J. Climate* **15**, 1141–1158. (doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2)
5. Fowler HJ, Ekström M. 2009 Multi-model ensemble estimates of climate change impacts and UK seasonal precipitation extremes. *Int. J. Climatol.* **29**, 385–416. (doi:10.1002/joc.1827)
6. Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M. 2010 Weight assignment in regional climate models. *Clim. Res.* **44**, 179–194. (doi:10.3354/cr00916)
7. Kjellström E, Boberg F, Castro M, Christensen JH, Nikulin G, Sánchez E. 2010 Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. *Clim. Res.* **44**, 135–150. (doi:10.3354/cr00932)
8. Lenderink G. 2010 Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations. *Clim. Res.* **44**, 151–166. (doi:10.3354/cr00946)
9. Kjellström E, Giorgi F. 2010 Introduction. *Clim. Res.* **44**, 117–119. (doi:10.3354/cr00976)
10. Déqué M, Somot S. 2010 Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments. *Clim. Res.* **44**, 195–209. (doi:10.3354/cr00866)
11. Weigel AP, Knutti R, Liniger MA, Appenzeller C. 2010 Risks of model weighting in multimodel climate projections. *J. Climate* **23**, 4175–4191. (doi:10.1175/2010JCLI3594.1)
12. Jun MR, Knutti R, Nychka DW. 2008 Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? *J. Am. Statist. Assoc.* **103**, 934–947. (doi:10.1198/016214507000001265)
13. Leith NA, Chandler RE. 2010 A framework for interpreting climate model outputs. *J. Roy. Stat. Soc. C* **59**, 279–296. (doi:10.1111/j.1467-9876.2009.00694.x)
14. Watterson IG. 2008 Calculation of probability density functions for temperature and precipitation change under global warming. *J. Geophys. Res.* **113**, D12106. (doi:10.1029/2007JD009254)
15. Tebaldi C, Smith RL, Nychka D, Mearns LO. 2005 Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *J. Climate* **18**, 1524–1540. (doi:10.1175/JCLI3363.1)
16. Greene AM, Goddard L, Lall U. 2006 Probabilistic multimodel regional temperature change projections. *J. Climate* **19**, 4326–4343. (doi:10.1175/JCLI3864.1)
17. Furrer R, Sain SR, Nychka D, Meehl GA. 2007 Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environ. Ecol. Stat.* **14**, 249–266. (doi:10.1007/s10651-007-0018-z)
18. Smith RL, Tebaldi C, Nychka D, Mearns LO. 2009 Bayesian modeling of uncertainty in ensembles of climate models. *J. Am. Statist. Assoc.* **104**, 97–116. (doi:10.1198/jasa.2009.0007)
19. Buser CM, Künsch HR, Lüthi D, Wild M, Schär C. 2009 Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Climate Dyn.* **33**, 849–868. (doi:10.1007/s00382-009-0588-6)
20. Tebaldi C, Sansó B. 2009 Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *J. Roy. Stat. Soc. A* **172**, 83–106. (doi:10.1111/j.1467-985X.2008.00545.x)
21. Buser CM, Künsch HR, Weber A. 2010 Biases and uncertainty in climate projections. *Scand. J. Stat.* **37**, 179–199. (doi:10.1111/j.1467-9469.2009.00686.x)
22. Kang EL, Cressie N, Sain SR. 2012 Combining outputs from the North American Regional Climate Change Assessment Program by using a Bayesian hierarchical model. *J. Roy. Stat. Soc. C* **61**, 291–313. (doi:10.1111/j.1467-9876.2011.01010.x)
23. Stainforth DA, Allen MR, Tredger ER, Smith LA. 2007 Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. R. Soc. A* **365**, 2145–2161. (doi:10.1098/rsta.2007.2074)
24. Rougier J. 2007 Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change* **81**, 247–264. (doi:10.1007/s10584-006-9156-9)
25. Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA. 2010 Challenges in combining projections from multiple climate models. *J. Climate* **23**, 2739–2758. (doi:10.1175/2009JCLI3361.1)
26. Annan JD, Hargreaves JC. 2010 Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.* **37**, L02703. (doi:10.1029/2009GL041994)
27. Knutti R, Abramowitz G, Collins M, Eyring V, Gleckler P, Hewitson B, Mearns L. 2010 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 (eds TF Stocker, D Qin, G-K Plattner, M Tignor, P Midgley). IPCC Working Group I Technical Support Unit. Bern, Switzerland: University of Bern.

28. Casella G, Berger RL. 1990 *Statistical inference*. Pacific Grove, CA: Duxbury Press.
29. Davison AC. 2003 *Statistical models*. Cambridge, UK: Cambridge University Press.
30. Rougier J, Goldstein M, House L. 2012 Second-order exchangeability analysis for multi-model ensembles. Technical report, no. 12:01, Department of Mathematics, University of Bristol, UK. See http://www.maths.bris.ac.uk/research/stats/reports/2012/.
31. Young GA, Smith RL. 2005 *Essentials of statistical inference*. Cambridge, UK: Cambridge University Press.
32. Goldstein M, Wooff D. 2007 *Bayes linear statistics: theory and methods*. Chichester, UK: Wiley.
33. Goldstein M, Rougier JC. 2009 Reified Bayesian modelling and inference for physical systems. *J. Stat. Plan. Inference* **139**, 1221–1239. (doi:10.1016/j.jspi.2008.07.019)
34. Maraun D *et al.* 2010 Precipitation downscaling under climate change—recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **48**, RG3003. (doi:10.1029/2009RG000314)