Second-order exchangeability analysis for multi-model ensembles

Jonathan Rougier*
Department of Mathematics
University of Bristol, UK

Michael Goldstein
Department of Mathematical Sciences
Durham University, UK

Leanna House
Department of Statistics
Virginia Tech, Blacksburg VA, USA

File `mme3.tex`, February 21, 2012

Abstract

The challenge of understanding complex systems often gives rise to a multiplicity of models. It is natural to consider whether the outputs of these models can be combined to produce a system prediction that is more informative than the output of any one of the models taken in isolation. And, in particular, to consider the relationship between the spread of model outputs and system uncertainty. We describe a statistical framework for such a combination, based on the exchangeability of the models, and their co-exchangeability with the system. We demonstrate the simplest implementation of our framework in the context of climate prediction. Throughout we work entirely in means and variances, to avoid the necessity of specifying higher-order quantities for which we often lack well-founded judgements.

*Corresponding author: Department of Mathematics, University Walk, Bristol, BS8 1TW, UK; email `j.c.rougier@bristol.ac.uk`.

1 Introduction

A complex system, like the climate, or a volcano, or an epidemic, or the economy, is inherently interesting to scientists. Consequently at any one time there will be a multiplicity of models of a given complex system, varying not just across research groups, but also, within research groups, across different conceptual formulations, and within each formulation, possibly across different numerical resolutions. When initially formulated, these models may well have been largely explanatory. But under pressure to harness science to the goal of effective policy-making, many of them are now being used predictively, to make statements about future system behaviour under different types of intervention (including no-intervention).

How this transition from explanatory to predictive models came about is an intriguing question, addressed by Oreskes (2000). Oreskes asserts that there is no necessary connection between science and prediction, and cautions earth scientists against accepting without question that they should predict. Her caution seemed prophetic: a more recent work, 'Useless Arithmetic: Why Environmental Scientists Can't Predict the Future' (Pilkey and Pilkey-Jarvis, 2007) documents many failures in environmental science. Writing as statisticians, we attribute such failures primarily to not accounting for the limitations of the models when assessing (or, unfortunately, not assessing) uncertainty about system behaviour. Not assessing uncertainty effectively leaves the door open to the misinterpretation or misuse of model predictions.

Scientists ought to be wary of complex system prediction, until they are able to quantify the effect of the limitations of their models on their predictions. This would almost certainly involve these scientists collaborating with statisticians, given that this type of uncertainty is very complicated to represent and to assess. Sansó *et al.* (2008) and Vernon *et al.* (2010) are two encouraging examples of this type of collaboration. Realistically, though, there are not enough applied statisticians to go around. Is there another way to assess system uncertainty? That is the subject of this paper.

An alternative to considering the models one-at-a-time, is to consider them collectively. If a collection of models of roughly the same quality can be assembled, then instead of each scientist working individually to assess the uncertainty of his model, the scientists collectively can assess the uncertainty properties of the ensemble, and perhaps fewer applied statisticians are required. Such 'multi-model ensembles' are now a important feature of climate prediction, as used by the assessment reports of the Intergovernmental Panel on Climate Change (IPCC). As this paper discusses, we statisticians have an important contribution to make, in

providing statistical models for ensembles and the complex systems they purport to represent. These statistical models have to account for the fact that the physical models are likely to be much more similar to each other than any of them is to the actual system. In other words, they cannot be treated as 'independent' measurements on the actual system.

Intuitively, one might imagine that the physical models in the ensemble might be 'independent' realisations of some 'representative model' for the ensemble, and that the system might be related to the individual models via the representative model. This intuition exactly anticipates the co-exchangeable statistical model that we present in this paper. Section 2 outlines our co-exchangeable model for a multi-model ensemble, and derives an important new result showing that a competitor model ('truth-plus-error') is in fact a restriction of our co-exchangeable model. Section 3 discusses the current situation in climate science, ahead of the next assessment report of the IPCC. The truth-plus-error model is hegemonic in this application, and we discuss the consequences of this, and of more complicated framework which permits conditional model-reweighting. Section 4 illustrates our co-exchangeable model with a simple application of climate prediction, and shows that when the ensemble is sufficiently large, useful predictions can be achieved with very limited judgements. It includes a simple and general approach to regularising large variances matrices, which is crucial in practice, when the number of model outputs can be two or more orders of magnitude larger than the number of models in the ensemble. Section 5 provides a brief summary.

Terminology. Throughout this paper we refer to the system models as 'simulators'. This is because the word 'model' is heavily overloaded, as a noun, a verb, and an adjective. Within 'statistical model' we include both the simulator outputs, *and* the judgements about how those relate to actual climate and climate observations. This includes an assessment of simulator quality. Therefore 'model criticism' is criticism of the whole statistical framework, not just the simulator. Statisticians have been using the phrase 'model criticism' in this way for decades (see, e.g., Box, 1980), but in the present context this requires a clear distinction between the computer code representing the physical model (i.e. the simulator) and the statistical model that encompasses it.

Also note that in this paper about judgements, 'we' denotes 'we, the authors' rather than 'you and I, dear reader', which J.M. Ziman termed the "diplomatic we" (quoted in footnote 67 of McIntyre, 1997).

2 The co-exchangeable model

Consider the three sets of operationally defined quantities: $\mathbb{X} := \{X^1, \ldots, X^m\}$, a collection of p outputs from each of m different simulators; Y the true values of the system that the simulators all purport to represent; and Z, a noisy and partial measurement of Y. In the usual arrangement for future prediction, X^i and Y would contain both historical and future components, and Z would be measurements on linear combinations of the historical components; p could easily be in the thousands. Our intention is to use the outputs of the simulators plus the observations to learn about the system, and hence we require a statistical model for $\mathbb{X}$, Y, and Z. It is standard to treat Z as conditionally independent of $\mathbb{X}$ given Y, and so the focus here on the joint statistical model of $\mathbb{X}$ and Y.

In this paper we use a second-order framework for our statistical model, namely the Bayes linear approach, e.g. as presented in Goldstein and Wooff (2007). In this approach expectation is primitive, and uncertainty about a collection of quantities such as $\{\mathbb{X}, Y\}$ is represented in terms of mean vectors and variance matrices. This collection can in principle be extended to indicator functions of any finite partition of the sample space, and hence be effectively equivalent to a fully probabilistic specification. But we can also proceed with a much coarser collection, suitable in situations where we feel unable to make the kinds of judgements necessary for a fully-probabilistic assessment. For complex systems, specifying means and variances is already a challenge, and in this paper we will not go any further than this. Where the assessment of uncertainty is a challenge, we prefer to make our means and variances explicit, allowing for discussion and refinement, and—one hopes—the gradual emergence of a consensus.

It is an interesting feature of the multivariate Gaussian distribution that its condition mean and variance expressions are identical in their form to the updated mean and variance expressions in the Bayes linear approach. Therefore readers who are uncomfortable working solely in terms of means and variances can simply adopt the Gaussian distribution throughout, and re-interpret updated means and variances as conditional means and variances. We stress, however, that this is not at all necessary, and that the Bayes linear approach has a coherence that is independent of Gaussian probabilistic assumptions.

The components of $\mathbb{X}$ are closely related to each other, and each of them is closely related to Y. To specify a statistical model that reflects these relationships requires the introduction of additional uncertain quantities that are *instrumental*. That is, they are not operationally defined; rather, they exist in order to help us to structure our judgements. The standard framework for this is exchangeability

(see, e.g., Schervish, 1995, chapter 1). Because we are adopting a second-order inferential framework, we will use the second-order exchangeability Representation Theorem of Goldstein (1986). An equivalent treatment would be in terms of second-order random effects, but we prefer to use exchangeability because it is rooted in judgements about operationally-defined quantities. All references to 'exchangeability' in this paper concern second-order exchangeability, with 'full exchangeability' being used for the fully probabilistic case.

2.1 The co-exhangeable model

Our statistical model for $\mathbb{X}$ and Y proceeds in two stages. First, it states that the simulator outputs in the ensemble are exchangeable. That is to say, each X^i has the same mean vector and variance matrix, and each pair $\{X^i, X^j\}$ has the same covariance matrix. Hence, according to the Representation Theorem, the collection of simulator outputs can be written

$$X^i = \mathcal{M}(X) \oplus \mathcal{R}^i(X) \qquad i = 1, \ldots, m, \tag{1a}$$

where the binary operator $\oplus$ is used to indicate the addition of components that are uncorrelated, $\mathcal{M}(X)$ is a common term, and the $\mathcal{R}^i(X)$ are residuals that have zero means, a common variance, and are uncorrelated across the simulators. Technically, this representation is formally correct only in the case in which our m simulators can be thought of as a subset of an infinitely exchangeable sequence, analogous to the fully probabilistic Representation Theorem of de Finetti (1937). In both cases, though, second-order and fully probabilistic, the infinitely exchangeable representation can be taken as a tractable approximation to finite exchangeability for sufficiently large m. It can be helpful to think of $\mathcal{M}(X)$ as the 'representative simulator' for the ensemble.

Second, our statistical model states that the system value Y respects exchangeability with the ensemble. That is to say, the covariance between Y and each X^i is the same. This implies that the relationship between the system value and the simulators can be written

$$Y = A\mathcal{M}(X) \oplus U \tag{1b}$$

where A is a known matrix, and U is uncorrelated with $\mathcal{M}(X)$ and the $\mathcal{R}^i(X)$; see Goldstein (1986) or Goldstein and Wooff (2007, section 7.1). We say that Y is 'co-exchangeable' with $\mathbb{X}$. The quantity U is the 'ensemble discrepancy', and represents errors that all simulators share.

The exchangeability of the simulators represents the judgement that the simula-

tor names are uninformative about simulator performance. Thus, it is a statement of ignorance about the simulators. The co-exchangeability of Y with $\mathbb{X}$ represents the judgement that all simulators are equally good (or bad) at simulating the system. This is a very natural follow-up judgement if the simulators themselves are represented as exchangeable. In fact, it is hard to imagine a defensible set of judgements in which $\mathbb{X}$ is exchangeable but Y is not co-exchangeable with $\mathbb{X}$, recalling that each X^i is trying to represent Y. In the absence of strong judgements, the default choice for A would be $A = I$, and the default choice for $E(U)$ would be $E(U) = \mathbf{0}$.

Crucially, in this model the joint collection $\{\mathbb{X}, Y\}$ is *not* exchangeable, except in the special case where $A = I$, $E(U) = \mathbf{0}$, and $\mathrm{Var}(U) = \mathrm{Var}(\mathcal{R}(X))$. This seems too strong a judgement for a complex system. Denote the ensemble mean as $\bar{X}$. The natural judgement for complex systems, allowing for common flaws in the simulators, would be that $\mathrm{Var}(X^i - \bar{X}) < \mathrm{Var}(Y - \bar{X})$, i.e. each of the simulators are more like the ensemble mean than the system is like the ensemble mean. This would imply $\mathrm{Var}(\mathcal{R}(X)) < \mathrm{Var}(U)$, taking $\bar{X} \approx \mathcal{M}(X)$ and $A = I$ for simplicity.

The co-exchangeable model implies the following variance and covariance properties for operationally-defined quantities:

$$\mathrm{Cov}(X^i, X^j) = \mathrm{Var}(\mathcal{M}(X)) \tag{2a}$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\mathcal{M}(X)) + m^{-1}\,\mathrm{Var}(\mathcal{R}(X)) \tag{2b}$$

$$\mathrm{Cov}(\bar{X}, Y) = \mathrm{Var}(\mathcal{M}(X))A^T \tag{2c}$$

$$\mathrm{Var}(Y) = A\,\mathrm{Var}(\mathcal{M}(X))A^T + \mathrm{Var}(U). \tag{2d}$$

Hence $\mathrm{Cov}(X^i, X^j)$ is non-negative definite symmetric, and $\mathrm{Cov}(\bar{X}, Y)$ is non-negative definite symmetric if $A = I$. The difference between the system value and the ensemble mean is $Y - \bar{X} = (A - I)\mathcal{M}(X) \oplus U \oplus m^{-1} \sum_{i=1}^{m} \mathcal{R}^i(X)$. In the default case where $A = I$, the variance simplifies to

$$\mathrm{Var}(Y - \bar{X}) = \mathrm{Var}(U) + m^{-1}\,\mathrm{Var}(\mathcal{R}(X)) \tag{3}$$

showing that the ensemble mean would not, in general, converge in mean square to the system value, owing to the presence of the ensemble discrepancy. In the stronger restriction in which $\{\mathbb{X}, Y\}$ are exchangeable, $Y - \bar{X}$ has mean zero and variance $(1 + m^{-1})\,\mathrm{Var}(\mathcal{R}(X))$. Even in this case the ensemble mean would not converge in mean square to the system value. There would always remain some uncertainty about the system value, no matter how large an ensemble it was pos-

sible to acquire. In the co-exchangeable model, the only way to get the ensemble mean to converge to the system value is to assert that there is *no* common error in the simulators, setting $\mathrm{Var}(U) = \mathbf{0}$.

2.2 The truth-plus-error model

As a demonstration that the co-exchangeable model is a very general statistical model for $\mathbb{X}$ and Y, we now show that it is compatible with what appears to be a quite different model. This is the 'truth-plus-error' model, which asserts that

$$X^i = Y \oplus B \oplus \mathcal{R}^i(X) \qquad i = 1, \ldots, m \tag{4a}$$

where B is a common 'bias' term that all simulators share, and $\mathcal{R}^i(X)$ is residual term as before, with mean zero and common variance. The etymology of the name is clear: each simulator tries to be the truth, but fails due to an error, one part of which is common, and one part of which is unique. The label 'truth-plus-error' has been adopted in climate science, discussed in more detail in section 3.

At first glance, the truth-plus-error model would appear to be incompatible with the co-exchangeable model of the previous subsection, because in that model Y is on the left of the equality and X^i is on the right (the right being the place where uncorrelated terms are added), while in this model X^i is on the left and Y is on the right.

To show the compatibility of the two models, start with the exchangeable representation for the simulators, (1a). Now consider writing the mean component of X^i as

$$\mathcal{M}(X) = Y \oplus B. \tag{4b}$$

Clearly (1a) and (4b) imply (4a). But (4b) also implies (1b), for a certain choice of A and U, namely

$$A' = \mathrm{Var}(Y)\,\mathrm{Var}(\mathcal{M}(X))^{-1} \quad \text{and} \quad U' = (I - A')\mathcal{M}(X) - B, \tag{4c}$$

where $\mathcal{M}(X)$ and U' are uncorrelated. Hence $A' \neq I$, but this is compensated by $E(U') = (I - A')\,E(Y) \neq \mathbf{0}$, if we make the default choice of $E(B) = \mathbf{0}$. (Or else the mean could be absorbed into B, and then U' could have zero mean.)

To derive this result, start by writing $Y = A'\mathcal{M}(X) + U'$, where necessarily $U' = (I - A')\mathcal{M}(X) - B$, in order that $Y = \mathcal{M}(X) - B$. This has the desired form of (1b), as long as we can choose A' such that $\mathcal{M}(X)$ and U' are uncorrelated. Solving $\mathrm{Cov}(\mathcal{M}(X), U') = \mathbf{0}$ implies that $A' = \mathrm{Var}(Y)\,\mathrm{Var}(\mathcal{M}(X))^{-1}$.

The really interesting question is whether the co-exchangeable model and the truth-plus-error model are equivalent, or whether the latter is a restriction on the former. If they are equivalent, then we may choose whichever form seems most natural for the specification of judgements. For the co-exchangeable model, the quantities that we can specify are $\mathrm{Var}(\mathcal{M}(X))$, A, $E(U)$ and $\mathrm{Var}(U)$, while for the truth-plus-error model they are $\mathrm{Var}(\mathcal{M}(X))$, $E(B)$ and $\mathrm{Var}(B)$. Enumerating these quantities shows immediately that the truth-plus-error model is a restriction of the co-exchangeable model, because it has less freedom in its specification, namely A is constrained.

So if You find Yourself contemplating the co-exchangeable model, yet unwilling to make a specific choice for A, not even the default choice $A = I$, then You may want to consider the restriction to the truth-plus-error model, in which case $A \neq I$ is set implicitly by Your choices for $\mathrm{Var}(\mathcal{M}(X))$ and $\mathrm{Var}(B)$. The key point is that the co-exchangeable model is the starting point, and selecting the truth-plus-error model represents a decision to impose a restriction on the co-exchangeable model, which would have to be justified in the context of the application.

2.3 Inference in the co-exchangeable model

To update our judgements about Y in the co-exchangeable model, we treat Y as sufficient for Z, as is standard. In our approach, this requires only that $\mathrm{Cov}(X^i, Z \mid Y) = \mathbf{0}$. Hence the update of Y by $\{\mathbb{X}, Z\}$ can proceed in two stages: first $\mathbb{X}$ is used to update Y, and then Z is used to update the updated Y.

Second, for the first update, note that $\bar{X}$ is Bayes linear sufficient for $\mathbb{X}$ when updating Y. This is because Y respects exchangeability with $\mathbb{X}$. Consequently, the update by $\mathbb{X}$ is invariant to permutations of its m components, and because the update is a linear combination of these outputs, the coefficients on each X^i must be identical, and no information is lost by replacing $\mathbb{X}$ with its arithmetic mean.

For the first stage, we can distinguish between applications where m is sufficiently large, and those where it is not. In the latter case, there is a fourth-order update that allows us to learn the $\mathrm{Var}(\mathcal{R}(X))$ from the ensemble, taking account of prior judgements (Goldstein and Wooff, 2007, ch. 8). In the former case, the ensemble variance matrix S provides a reasonable estimate of $\mathrm{Var}(\mathcal{R}(X))$, possibly after regularisation (discussed in section 4.3). Likewise, we can perform a second-order update for the mean of Y, or we can adopt the simpler practice of using the observed ensemble mean, $\bar{x}$. We adopt both of these practices here, for simplicity, and because they are consistent with our application in section 4.

Hence $\mathrm{Var}(\mathcal{R}(X)) = S$ and $E_{\bar{X}}(Y) = A\bar{x}$.

For the updated variance of Y, recollect that the updated variance in a Bayes linear analysis has the same form as the conditional variance of a Gaussian collection. Hence, plugging in from (2),

$$\begin{aligned}\mathrm{Var}_{\bar{X}}(Y) = A\,\mathrm{Var}(\mathcal{M}(X))A^T + \mathrm{Var}(U) - \\ A\,\mathrm{Var}(\mathcal{M}(X))\big\{\,\mathrm{Var}(\mathcal{M}(X)) + m^{-1}S\big\}^{\dagger}\,\mathrm{Var}(\mathcal{M}(X))A^T,\end{aligned} \tag{5}$$

where the superscript '†' indicates the Moore-Penrose inverse. An interesting feature of this update is that it is insensitive to the value of $\mathrm{Var}(\mathcal{M}(X))$, provided that $\mathrm{Var}(\mathcal{M}(X)) \gg m^{-1}S$, i.e. if the ensemble is sufficiently large. In this case $\mathrm{Var}_{\bar{X}}(Y) \approx \mathrm{Var}(U)$. Effectively, if the ensemble is sufficiently large then it constrains $\mathcal{M}(X)$ to such a degree that its variance becomes negligible. But it does not constrain U at all (which is uncorrelated with $\mathbb{X}$), and consequently the updated variance for $Y = A\mathcal{M}(X)+U$ is approximately $\mathrm{Var}(U)$. The same argument justifies setting $E_{\bar{X}}(Y) = A\bar{x}$.

The second stage update is based on a standard linear model for Z, namely

$$Z = H\,Y \oplus W \tag{6}$$

where H is a known incidence matrix and W is representation and measurement error, which we will assume has mean zero and variance D. This variance is often treated as diagonal, due to the dominance of the measurement error. After the first stage we have

$$E_{\bar{X}}(Z) = H\mathbb{E} \quad \text{and} \quad \mathrm{Var}_{\bar{X}}(Z) = H\mathbb{V}H^T + D, \tag{7}$$

where $\mathbb{E} = E_{\bar{X}}(Y) \approx A\bar{x}$ and $\mathbb{V} = \mathrm{Var}_{\bar{X}}(Y) \approx \mathrm{Var}(U)$. This 'update by noisy linear combinations' has a tractable computational form provided that all variances are non-singular, found by applying the Sherman-Woodbury-Morrison formula:

$$\mathrm{Var}_{\bar{X},Z}(Y) = \big\{\mathbb{V}^{-1} + H^T D^{-1} H\big\}^{-1} \tag{8a}$$

$$E_{\bar{X},Z}(Y) = \mathbb{E} + \mathrm{Var}_{\bar{X},Z}(Y)H^T D^{-1}\big(Z - H\mathbb{E}\big). \tag{8b}$$

There is an interesting coda to this calculation in the case of the truth-plus-error model. We have shown that if A is specified directly the value of $\mathrm{Var}(\mathcal{M}(X))$ drops out of the update for Y if the number of simulators is sufficiently large. But in the truth-plus-error model $\mathrm{Var}(\mathcal{M}(X))$ remains, through its role in setting A'.

This affects both the mean and variance of Y. So the irony of the truth-plus-error model is that it seems to remove the necessity of specifying A, but in fact this comes at the price of having to specify $\mathrm{Var}(\mathcal{M}(X))$ even when the number of simulators is large.

2.4 Model criticism

Our judgements about the instrumental quantities in the co-exchangeable model can be assessed by their implications for the operationally-defined quantities, as shown in (2). But where our judgements about the operationally-defined quantities are themselves quite limited, we seek the reassurance of diagnostic evaluation. There are two levels of diagnostic evaluation: of the exchangeability of the ensemble, and of the relationship between the ensemble and the measurements.

The first stage is easily implemented. Under the judgement of exchangeability, we can use $m-1$ simulators to predict the output of the mth, for each of the m simulators. Smith *et al.* (2009, section 4.1) provides a fully-probabilistic implementation of this cross-validation approach in climate science. For a second-order approach we compare each X^i with its predictive mean vector and variance matrix, $E_{\bar{X}^{-i}}(X^i)$ and $\mathrm{Var}_{\bar{X}^{-i}}(X^i)$. As above, the updated mean and variance are insensitive to the value of $\mathrm{Var}(\mathcal{M}(X))$ if the ensemble is sufficiently large, in which case $E_{\bar{X}^{-i}}(X^i) \approx \bar{x}^{-i}$ and $\mathrm{Var}_{\bar{X}^{-i}}(X^i) \approx S^{(-i)}$.

For the second stage diagnostic we compare the actual value of Z against the predictive mean and variance, as given in (7). This is a 'full model diagnostic': it incorporates our judgement of exchangeability for $\mathbb{X}$, the co-exchangeability of Y with $\mathbb{X}$, and the choices we have made for the instrumental quantities.

3 The situation in climate science

Quantifying climate uncertainty is now a pressing scientific concern, given the crucial role that uncertainty plays in the evaluation of different strategies for climate change mitigation and adaptation. Under the auspices of the Intergovernmental Panel on Climate Change (IPCC), research groups across the world have run climate projections under specific future scenarios, each using their own climate simulators; the outputs of these runs are referred to here, collectively, as the Multi-Model Ensemble (MME). The results have been central to the analysis in the IPCC Fourth Assessment Report (AR4), and the updated MME will play a similar role in the Fifth Assessment Report.

In ch. 10 of AR4 Working Group 1 (Meehl *et al.*, 2007), section 10.5 addresses

the issue of "Quantifying the range of climate change projections". It is recognised that the range of climate uncertainty can only be partially quantified by the range of the MME, due to "common systematic biases" (Meehl *et al.*, 2007, p. 797). No attempt is made, formally, to account for these common biases in assessing ranges for global mean temperature under different future scenarios. Instead, the authors propose a range of -40% to $+60\%$ of the ensemble mean, based on expert judgement that synthesises evidence from a wide variety of studies (see, e.g, Meehl *et al.*, 2007, Figure 10.29, p. 809).

Unsurprisingly, given the topicality of the area and the richness of the dataset, statisticians have been involved in providing climate predictions that combine the MME and climate observations, and that provide a more sophisticated quantification of uncertainty. Tebaldi and Knutti (2007) and Knutti *et al.* (2010b) provide reviews, and the IPCC have also produced a recent Good Practice Guidance Paper (Knutti *et al.*, 2010a). We start by outlining two current approaches.

Tebaldi *et al.* (2010) describe the standard statistical model for the climate MME (see also, among others, Tebaldi *et al.*, 2005; Furrer *et al.*, 2007; Berliner and Kim, 2008; Tebaldi and Sansó, 2009; Buser *et al.*, 2009; Smith *et al.*, 2009). We will refer to this as the *Tebaldi model*. It is important to understand the origins of this model. When simulators in the MME were first combined into a single point prediction for climate, it was felt sensible to weight them according to two factors: their ability to reproduce climate observations, and their closeness to the mean of the ensemble. This was termed *Reliability Ensemble Averaging* (Giorgi and Mearns, 2002). Tebaldi *et al.* (2005) showed that this can be achieved within a formal statistical model for $\mathbb{X}$, Y, and Z of the general form

$$X^i = Y \oplus \lambda_i^{-1/2} R^i \qquad i = 1, \ldots, m \tag{9a}$$

$$Z = HY \oplus \lambda_0^{-1/2} W, \tag{9b}$$

where the R^i are mean zero, uncorrelated, with a common variance. The λ's are simulator-specific precisions. Conditional on these λ's, the updated mean for actual climate is the weighted average of $X^1, \ldots, X^m, Z$, where the X^i weight is λ_i, and the Z-weight is λ_0 for components of Y that are measured, and zero otherwise.

Tebaldi *et al.* (2010) use a fully probabilistic model, rather than a second-order model. This has the standard conjugate form: X, Y, and Z are conditionally Gaussian given some precision parameters, and the precision parameters have tractable Gamma distributions. The λ_i are treated as *a priori* fully exchangeable, and therefore the X^i are *a priori* fully exchangeable as well. However, after conditioning on

$\mathbb{X}$ and Z, the posterior mean of λ_i is decreasing in both $(HX^i - Z)^2$ and $(X^i - \bar{X})^2$, leading to weightings that favour simulators which can replicate historical climate measurements, and which are close to the ensemble mean. (This is a heuristic explanation adapted to our notation.)

A natural generalisation of (9) that is commonly adopted is to include an independent bias term in the X^i equation, with a common component and a simulator-specific component (Tebaldi *et al.*, 2010). The simulator-specific component can be pushed into R^i, and the result is

$$X^i = Y \oplus B \oplus \lambda_i^{-1} R^i \qquad i = 1, \ldots, m, \tag{10}$$

which is almost the truth-plus-error model, the only difference being the λ's. The presence of uncertain and simulator-specific λ_i is crucial in achieving a differential weighting across the simulators. It impairs the tractability of the co-exchangeable model, because it stops the ensemble mean from being Bayes linear sufficient. However, for fixed λ's the Bayes linear calculation is still straightforward, if tedious, and the λ's themselves can be updated in a fourth-order analysis if required (Goldstein and Wooff, 2007, ch. 8). The fully-probabilistic model is fitted with a Gibbs sampler, including Metropolis-Hastings steps for variance parameters.

Chandler (2010) suggests an alternative model (see also Leith and Chandler, 2010). This model is (in our notation)

$$X^i = X^0 \oplus V^i \qquad i = 1, \ldots, m \tag{11a}$$

$$Y = X^0 \oplus U \tag{11b}$$

$$Z = HY \oplus W. \tag{11c}$$

where the V's are mean zero and uncorrelated, but where $\mathrm{Var}(V^i)$ might differ across simulators. This is close to the co-exchangeable model with $A = I$. Chandler treats all variances as known, and in a *tour de force* of matrix algebra is able to derive an explicit second-order updated mean and variance for Y. In the special case where $\mathrm{Var}(V^i)$ is the same for all simulators this becomes a special case of our co-exchangeable model, and the update depends only on the ensemble mean and not on the individual simulator outputs, as we explained in section 2.3. A special case of this model is the hypothesis advanced by Annan and Hargreaves (2010), that $\{\mathbb{X}, Y\}$ are exchangeable, although they say "statistically indistinguishable".

We assess these two models in the light of our own co-exchangeable model. First, we note that our preference when modelling complex systems such as cli-

mate is to stick with second-order representations of uncertainty. We have absolutely nothing against fully probabilistic models, in situations where judgements are available to support the specification of higher moments. Climate scientists are familiar with means and variances, and with the second-order updating formulae, partly because of the widespread use of the Kalman filter for data assimilation in meteorology and reconstruction (see, e.g., Talagrand, 1997). But representing uncertainty as variance matrices is already a challenge, even for statisticians. We advocate explicit statements about prior variances, so that these judgements may be debated and refined, and not delegated to additional levels of a hierarchical statistical model. We are suspicious of higher-order judgements about complex systems made purely for the purposes of tractability (Rougier, 2006).

Second, we note that, the λ's apart, the Tebaldi model is a truth-plus-error model. The IPCC Good Practice Guidance Paper (Knutti *et al.*, 2010a, p. 4) makes a mistaken distinction between the truth-plus-error model and the co-exchangeable model, which we are here able to correct: the truth-plus-error model is a restriction of the co-exchangeable model, as shown in section 2.2. As already explained, the truth-plus-error model removes the need to make an explicit choice for A. But if specifying A is perceived to be a problem, then considerations of simplicity and transparency lead us to favour the default choice of $Y = \mathcal{M}(X) + U$ where $A = I$ and $E(U) = \mathbf{0}$, rather than $Y = A'\mathcal{M}(X) + U'$ where $A' \neq I$ and $E(U') \neq \mathbf{0}$, which is the truth-plus-error model, where both A' and $E(U')$ will depend on the choice of $\mathrm{Var}(\mathcal{M}(X))$. As explained in section 2.3, another benefit of the co-exchangeable model with the default choice is that the value of $\mathrm{Var}(\mathcal{M}(X))$ drops out of the inference (m being sufficiently large in climate science, as we show in section 4.4), but not in the truth-plus-error model.

Third, both the Tebaldi and Chandler models have the potential to attach different weights to the simulators in the MME, either through prior considerations, or as a consequence of variance learning in the posterior. Whilst our co-exchangeable model could also be generalised in the same way, we would like to make a case for sticking with exchangeability. Obviously, if judgements are strong enough to quantify differences across simulators then these should be incorporated. But how realistic is this in climate science?

Consider the simplest case, of comparing two climate simulators from the same research group: the established simulator which has been around for nearly a decade, and the new simulator which has been developed over the last couple of years. One can point to improvements in the modules: the new simulator may have a fancier ocean module, with better physics and a higher resolution. But

climate simulators are very complex artifacts, with many thousands of lines of code and hundreds of interacting user-specified parameters. There is a lot to be said for a simulator that has been operational for a decade, given that there is no formal way to test the correctness and stability of such a complicated code, only the accumulation of experience.

So thinking seriously about how the two simulators are different in their representations of actual climate is a demanding and time-consuming task, even in this simple case which ignores scientific and cultural differences between research groups. This is especially true when dealing with large collections of outputs, such global temperature at 5° resolution ($\sim$2500 outputs), which will be our application in section 4. Climate scientists may choose to adopt a simple baseline statistical model for the purposes of climate inference, reserving the option to upgrade it in due course if resources allow. Exchangeability is one such baseline model.

Exactly the same attitude is taken in the climate simulator itself. For example, in prescribing a grossly simplified vegetation scheme. Obviously it would be better to develop and implement a responsive and dynamical scheme, but that would introduce more user-specified parameters and cost more CPU cycles. The fact that a climate simulator has a prescribed vegetation scheme does not render the simulator uninformative in the minds of climate scientists. In the same way, inferences based on an initial treatment of exchangeability are not invalidated by knowledge about differences between simulators.

4 Application: global surface temperatures

This application uses the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (`http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php`). We focus on a very small part of the full dataset, namely boreal winter (i.e. December to February, or DJF) mean surface temperature for the period 1980–1999, aggregated to $5° \times 5°$ gridcells. Each simulator-output is a 2448-vector, representing a 72×34 collection of gridcell values covering the longitude interval $[-180, 180]$ and the latitude interval $[-85, 85]$. Jun *et al.* (2008b) provide details of this dataset, and a comparison between the simulators (see also Jun *et al.*, 2008a).

Throughout this section we will use the simplest implementation of our co-exchangeable model. We have a sufficiently large ensemble that the value of $\mathrm{Var}(\mathcal{M}(X))$ drops out, to a good approximation. We adopt the default choices of $A = I$ and $E(U) = \mathbf{0}$. And we use simple approaches for estimating $\mathrm{Var}(\mathcal{R}(X))$

from the ensemble, and for specifying $\mathrm{Var}(U)$. All of these implementation decisions may be generalised where additional judgements permit. We make them here partly as a reflection of our own limited judgements in this application, and also from an academic interest in seeing how far one can go with minimal judgements, beyond the initial selection of the co-exchangeable model.

The data and our code for the calculations in this application is available in the Supplementary Online Materials (SOM), written in the statistical computing environment R (R Development Core Team, 2004). The maps should be viewed in colour (as explained immediately below), and are available in colour in the SOM.

4.1 Observations

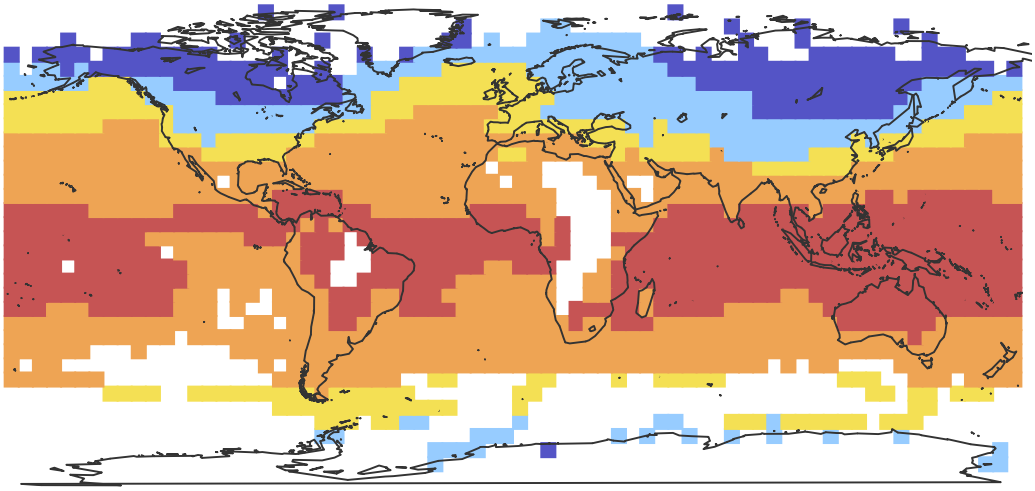
We use the same observations as Jun *et al.* (2008b), shown in Panel (a) of Figure 1, which have been interpolated onto the same grid as the simulator-outputs. The choice of colour-scale for such a map is subtle. We are experimenting in this paper with the use of hue for the mean field and saturation for the standard deviation field, following the recommendation and colour-scheme of Kaye *et al.* (2011); more details are given in Figure 2. This limits the resolution of either field, in our case to five hues for the mean and four saturation levels for the standard deviation (implying up to sixteen different colours on the map). We are not yet convinced of the advisability of representing both mean and standard deviation on a single map, but it is certainly worth a try, given that this would be an excellent way to ensure that uncertainty information does not become detached from point reconstructions.

Figure 1a shows that many observations are missing (759), particularly around the poles. Because of this, Jun *et al.* restricted their analysis to the latitude range 45°S to 72°N, and then interpolated the missing values inside this range. In our treatment we allow explicitly for the observations being only a subset of the quantities of interest—the gridcells with missing observations will play the role of the 'predictive' quantities. In other words, we will predict surface temperature for every point on the grid, even though our observations only make up a subset of these. Referring back to the observation equation, (6), H is implicit in Figure 1a, and we will treat the components of W as uncorrelated, with mean zero and standard deviation 0.5 °C.

4.2 Choice of simulators

Our first statistical choice is which subset of the 20 simulators in the MME we treat as exchangeable. Recall that judgements of exchangeability are judgements

(a) Observations



(b) Updated temperature

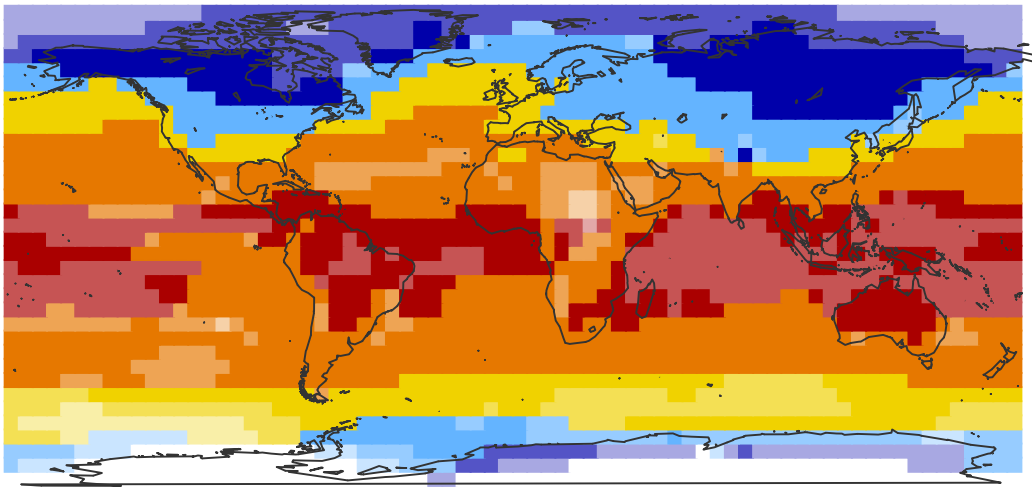


Figure 1: Surface temperature, 1980–1999, boreal winter (DJF months), aggregated to 5° gridcells. (a) Observations (white = missing), (b) Updated using both the multi-model ensemble and the observations. The colour scheme represents both the mean field and the standard deviation field. See Figure 2 for details.

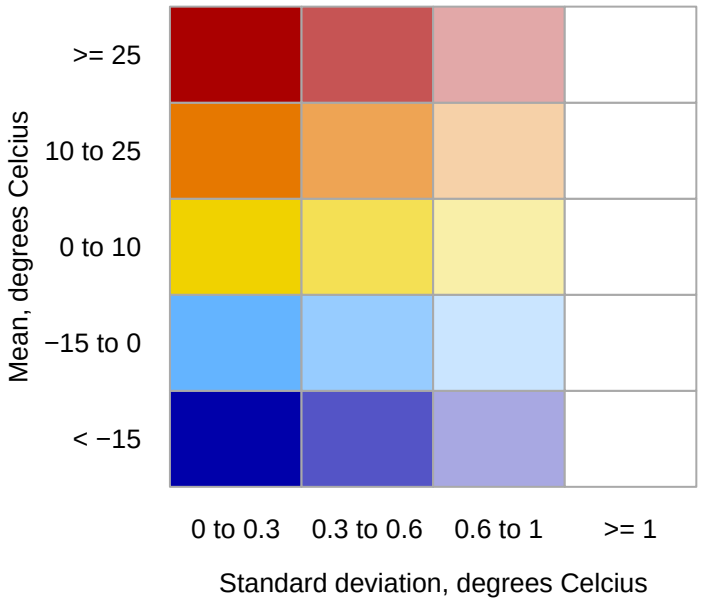


Figure 2: Legend for the Figure 1. Hue denotes the value of the mean field and saturation the value of the standard deviation field, following the recommendation and colour-scheme of Kaye *et al.* (2011), in their response to the discussion by Rougier.

of ignorance; to say that the components of $\{\mathbb{X}\}$ are exchangeable is to say that the indices $i = 1, \ldots, m$ are immaterial. In a climate MME these indices represent simulator names. To a climate scientist, these names contain a weath of information, about the group that constructed the simulator, and the version of the code. Thus they are not exchangeable.

However, without exchangeability this is an much more awkward inference, requiring potentially different judgements for each simulator, and without the benefit of the sufficiency of the ensemble mean. Our compromise is to work with a subset of the MME that we judge 'somewhat exchangeable'. This real or adopted ignorance is designed to get our inference up and running, by dramatically simplifying the statistical modelling, as discussed in section 3.

Given the incremental way in which climate simulators are constructed, and the large influence of scientific modelling judgements, outputs of simulators from the same research group will typically be more like each other than like other MME members, and so we will retain only one simulator from each research group. In each case, we keep the simulator which we judge is most like `UKMO-HadCM3`, which we take as our benchmark. Like Jun *et al.* we exclude the simulator from the Beijing Climate Center (BCC-CM1) on *a priori* grounds. This gives us 14

simulators which we will treat as exchangeable:

$$\mathcal{S} := \big\{\texttt{CGCM3.1}, \texttt{CNRM-CM3}, \texttt{CSIRO-Mk3.0}, \texttt{GFDL-CM2.0}, \texttt{GISS-AOM}, \texttt{FGOALS-g1.0}, \texttt{INM-CM3.0}, \texttt{IPSL-CM4}, \texttt{MIROC3.2}, \texttt{ECHO-G}, \texttt{ECHAM5/MPI-OM}, \texttt{MRI-CGCM2.3.2}, \texttt{PCM}, \texttt{UKMO-HadCM3}\big\}. \quad (12)$$

4.3 Judgements about the MME

Next we consider the exchangeable representation of our MME, which we continue to label $\mathbb{X}$, but which has been reduced to $\mathbb{X} := \{X^i : i \in \mathcal{S}\}$ for the set $\mathcal{S}$ given above; we continue to use m for the number of simulators in $\mathbb{X}$. We will use the ensemble mean $\bar{x}$ to estimate $E(\mathcal{M}(X))$, and the ensemble variance S to estimate $\mathrm{Var}(\mathcal{R}(X))$. However, there is a catch: for a large set of outputs and a small number of simulators, the estimated variance S will be singular. We do not believe that all but a few linear combinations of $\mathcal{R}^i(X)$ will have zero variance, and therefore we must regularise our estimate in some way.

In general this regularisation takes the form of combining our prior judgements about $\mathrm{Var}(\mathcal{R}(X))$ with the information in the ensemble, namely S. In the Bayes linear approach this would be a fourth-order update, as discussed in Goldstein and Wooff (2007), notably sections 8.10–8.13. Here we provide a simplified approach that is consistent with our limited judgements about $\mathcal{R}^i(X)$. We want to achieve two goals: to make the variance full rank and to smooth out the covariance structure. We would like to preserve the trace of S in our regularised version, as we have no reason to think that the individual variances are systematically mis-represented. This suggests using a regularisation of the form

$$S^{\mathrm{reg}} = (1-\alpha)GRG + \alpha S \quad (13a)$$

where R is a spatially-smooth correlation matrix, G is a diagonal matrix of the standard deviations of S, and $\alpha \in [0,1]$. One suggestion in Goldstein and Wooff (2007, section 8.11) is that α can be set by valuing the information in R according to the number of ensemble members it is equivalent to, say m_0. Then $\alpha = m/(m_0+m)$. Another approach, which we propose here, is to make α a function of R and S. S has $m-1$ non-zero eigenvalues: denote the eigenvectors of these eigenvalues as Γ_1, which has dimension $p \times (m-1)$, where p is the number of simulator outputs. Then setting

$$\alpha = \mathrm{tr}(\Gamma_1^T R \Gamma_1)/p \quad (13b)$$

will upweight S relative to GRG when the linear combinations in Γ_1 have large

variances in R, i.e. they are spatially smooth in R. So if S is spatially smooth already we do not tamper with it, but if it is not smooth we average it with something that is. Note that the regularised estimator is unbiased for $m \geq p + 1$, and consistent when the ensemble is generated from a matrix Normal distribution. This approach to regularisation is similar to the 'semi-adjusted' approach of Goldstein and Wooff (2007, section 8.13). Other more classical approaches to regularisation, such as banding or tapering (see, e.g., Bickel and Levina, 2008), are not appropriate here due to the possibility of teleconnections between well-separated regions, which is a feature of the climate system, and of climate simulators.

To choose R we specify an isotropic correlation matrix, which has a single parameter (the correlation length). This correlation matrix needs to account for the geometry of the surface of the earth. We express the correlation between two locations separated by an angle θ by

$$\psi(\theta) = \varphi\left(2r \sin\left(\frac{\theta}{2}\right)\right) \qquad \theta \in [0, \pi] \tag{14}$$

where r is the radius of the Earth (6378 km), and φ is a correlation function; see, e.g., Gneiting (1999), or the discussion in Furrer *et al.* (2007). Given the longitude and latitude of the two locations, the angle between them can be computed using the Haversine formula (see, e.g., Sinnott, 1984). For $\varphi(\cdot)$ we use a Matérn correlation function with shape $\nu = 5/2$, for which

$$\varphi(d; \ell) = \left(1 + \frac{\sqrt{5}d}{\ell} + \frac{5d^2}{3\ell^2}\right) \exp\left(\frac{-\sqrt{5}d}{\ell}\right) \tag{15}$$

where ℓ is a scale parameter that sets the correlation length (see, e.g., Rasmussen and Williams, 2006, ch. 4). The shape parameter ν controls the smoothness of the sample paths; the value $\nu = 5/2$ implies that the sample paths are twice differentiable in mean square. We concur with Rasmussen and Williams (p. 85), in viewing this choice of ν as a reasonable and tractable compromise between $\nu = 3/2$, which is quite rough, and $\nu = 7/2$, which is hard to distinguish from the squared exponential, a correlation function which we regard as too smooth (being infinitely differentiable in mean square).

We choose a correlation length of one sixth, defined to be the proportion of the earth's circumference at which the correlation falls to 0.05, which implies that $\ell = 2410$ km. For us, this is 'continental scale': we do not expect there to be a lot of smoothness in $\mathcal{R}^i(X)$ on larger scales than this, due to land/ocean contrasts, but in a more sophisticated analysis we might include explicit large-scale teleconnections,

such as those from the El Niño Southern Oscillation.

Referring back to (13), this choice of R gives $\alpha = 0.25$ for our MME. This values R at $m_0 \approx 40$ ensemble members, which is larger than we would like, but not surprising given the huge discrepancy between the number of simulators and the number of simulator outputs. There is no avoiding the regularisation of S if we want to work at this level of spatial resolution, and so we pay close attention to the leave-one-out diagnostic assessment of this regularisation, described in section 2.4. Figure 3 shows the marginal and joint diagnostics for simulators five to eight, each based on a comparison of X^i with the mean $\bar{x}^{-i}$ and variance matrix $S^{(-i)}$. The marginal diagnostics plot the standardised prediction errors, $(X^i_j - \bar{x}^{-i}_j)/\sqrt{S^{(-i)}_{jj}}$, for $j = 1, \ldots, p$. The joint diagnostics show a smoothed histogram of the standardised prediction errors of all of the principle components of $S^{(-i)}$ that have variance of at least $(0.5\,^\circ\mathrm{C})^2$. That is, if $S^{(-i)} = \Gamma\Lambda\Gamma^T$, then the histogram shows $(X^i - \bar{x}^{-i})^T\gamma_j/\sqrt{\lambda_j}$ for all those j for which $\lambda_j \geq (0.5\,^\circ\mathrm{C})^2$.

Generally, the marginal diagnostics look fine, and the joint diagnostics look acceptable, given that it is very difficult to get covariances of large collections of highly-related quantities right (recollect also that $S^{(-i)}$ is only approximately the predictive variance). The exception is the simulator `FGOALS-g1.0`, which is quite different from the other simulators in the MME, as shown in Figure 3. Therefore we excluded it from further analysis. Jun *et al.* (2008b) also noted that this simulator had unusually large biases. A well-informed climate modeller may well have excluded this simulator *a priori*.

4.4 Linking the simulators to actual climate

We will set $A = I$ and $E(U) = \mathbf{0}$ in (1b), i.e. make the default choices. We have used the ensemble estimates $\bar{x}$ and S for $E(\mathcal{M}(X))$ and $\mathrm{Var}(\mathcal{R}(X))$, the latter after regularisation. This only leaves $\mathrm{Var}(U)$ to be assessed.

We bootstrap our way to an assessment of $\mathrm{Var}(U)$ by considering it in relation to S. First, we consider the natural judgement that when the simulators disagree on an output component, the ensemble mean is a less-reliable predictor for that component. We showed in (3) that $\mathrm{Var}(Y - \bar{X}) = \mathrm{Var}(U) + m^{-1}S$, for which the diagonal is increasing in the diagonal of S, and so this effect is 'built in' to the co-exchangeable model. However, it diminishes rapidly as the ensemble size increases, and for this reason we may want to augment the built-in effect by making $\mathrm{Var}(U)$ proportional to S. We have already argued that $\mathrm{Var}(U) = S$ is too strong for the current generation of climate simulators, because this would assert, in conjunction with our other choices, that $\{\mathbb{X}, Y\}$ was exchangeable. Therefore we propose to

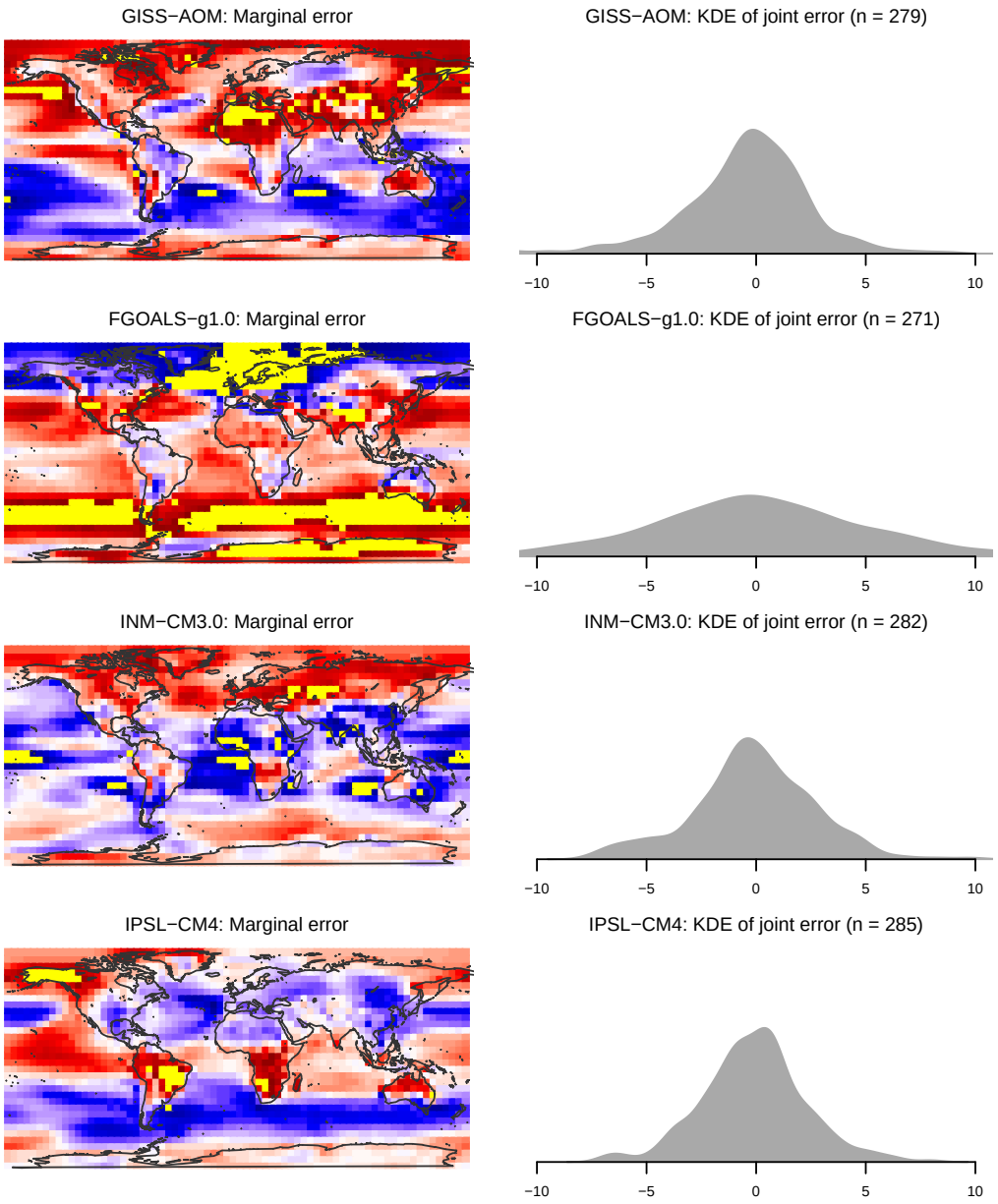


Figure 3: Diagnostic evaluation of the exchangeability of our ensemble, and our approach for regularising the estimated variance of $\mathcal{R}^i(X)$. Showing just simulators five to eight in (12). The left-hand panels show standardised marginal prediction errors on a scale from -3 (deep blue) to 3 (deep red), with more extreme values than this shown in yellow. The right-hand panels show the smoothed histogram of the standardised prediction errors of those principal components that have variance of at least $(0.5\,^{\circ}\mathrm{C})^2$. Figures for all simulators are available in the SOM.

model $\mathrm{Var}(U)$ as

$$\mathrm{Var}(U) = V_0 + \kappa_U^2 \, S \tag{16}$$

for some variance matrix V_0 and some scalar $\kappa_U > 1$, which we specify. The larger is κ_U, the less informative is the ensemble mean for actual climate. In a more general treatment we could treat κ_U as a diagonal matrix and allow it to vary component-by-component, for example by latitude in our application, or by type of variable, if there is more than one type. Experienced climate scientists may feel confident about expressing such judgements, but we are content to use a single value for all gridcells.

The presence of V_0 in (16) is to ensure that in our simple model for $\mathrm{Var}(U)$ we do not treat complete agreement among the climate simulators as synonymous with no ensemble discrepancy. We achieve an equivalent effect by lower-bounding each individual variance in S at $(1\,^\circ\mathrm{C})^2$, writing the lower-bounded variance as $\tilde{S}$, so that $\mathrm{Var}(U) = \kappa_U^2 \, \tilde{S}$. In our application most of the variances in S are much larger than $(1\,^\circ\mathrm{C})^2$, and so $\tilde{S} \approx S$.

We will use a similar device for $\mathrm{Var}(\mathcal{M}(X))$, but only for formal purposes, to establish that this variance can be neglected in our application. If we write $\mathrm{Var}(\mathcal{M}(X)) = \kappa_M^2 \, \tilde{S}$, then we can translate our choice of κ_M into a correlation, because

$$\mathrm{Corr}(X_j^i, X_j^{i'}) = \frac{\kappa_M^2}{\kappa_M^2 + 1} \tag{17}$$

where $i \neq i'$ and the subscript j indicates the jth component. Climate simulators are carefully tuned to get global and hemispherical temperatures about right, and so a correlation greater than 0.5 seems appropriate. This implies $\kappa_M > 1$.

With this simplification for $\mathrm{Var}(U)$, it is straightforward to compute the first stage update from subsection 2.3. The updated variance of Y is

$$\mathrm{Var}_{\bar{X}}(Y) = \left(\kappa_U^2 + \frac{\kappa_M^2 m^{-1}}{\kappa_M^2 + m^{-1}} \right) \tilde{S} \approx \kappa_U^2 \, \tilde{S} \tag{18}$$

where the approximation holds providing that $\kappa_M^2, \kappa_U^2 \gg m^{-1}$; it certainly holds in our case of $\kappa_M > 1$, $\kappa_U > 1$, and $m = 13$.

This derivation shows the conditions under which the updated variance of actual climate is approximately proportional to the ensemble variance (after regularisation and lower-bounding). But it does not justify setting $\kappa_U = 1$; as we have already argued, we will choose $\kappa_U > 1$.

4.5 Including the observations

Finally, we include the observations, Z, for which we have a known incidence matrix H and an error variance matrix $D = \text{diag}\{(0.5\,^\circ\text{C})^2\}$. We can critique our statistical modelling choices though the standardised prediction errors for Z, which have adjusted mean and variance given in (7). This is a 'whole model' diagnostic, because in order to get to this predictive mean and variance we have had to use the co-exchangeable model, and make the additional statistical judgements described in the previous two subsections. At the moment, κ_U remains unspecified: we can make an *a priori* assessment of it, or we can treat it as a tuning parameter for improving the diagnostics, which is a simple form of variance learning.

Figure 4 shows the diagnostic assessment, both the marginal standardised prediction errors, and the principal component standardised prediction errors, for four different values for κ_U: 0.75, 1, 1.25, and 1.5. Our favoured choice is 1.25, balancing our *a priori* preference for $\kappa_U > 1$ with the information from the diagnostic assessment.

Finally, Panel (b) of Figure 1 shows the updated mean and standard deviation of actual climate, using the information from both the ensemble and the observations, computed using (8). Where there were observations, the standard deviation has been brought down below the measurement error of 0.5 °C, as would be expected. There is more uncertainty in the polar regions, reflecting both the lower level of agreement between the climate simulators in those regions, and the scarcity of observations. Only in Antarctica is the updated standard deviation larger than 1 °C.

5 Summary and conclusion

How much judgement is required when combining the outputs from an ensemble of computer simulator runs, plus partial and noisy measurements, to make a statistical prediction of the behaviour of a complex system? We have argued and shown that if the ensemble is sufficiently large, and we have (or are content to adopt) a reasonable amount of ignorance, then one scalar is (almost) all that is required: κ_U, which quantifies which is closer to the ensemble mean, an individual simulator output or the actual system values (with $\kappa_U = 1$ indicating that they are equally close).

This conclusion is based on the adoption of our second-order co-exchangeable statistical model for the ensemble and the system, given in (1). This model is a statement of ignorance about the simulators, which might require us to be selective

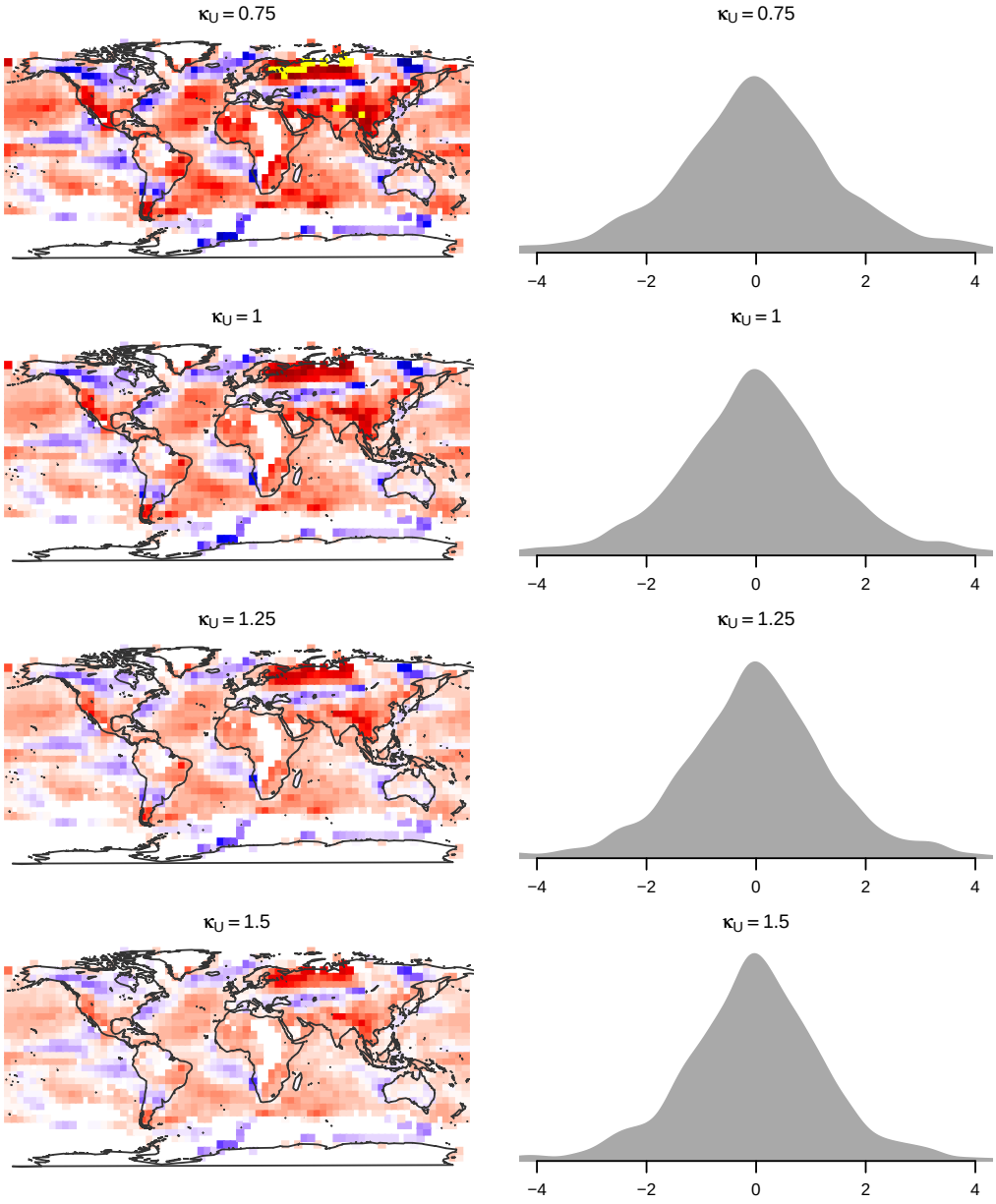


Figure 4: Diagnostic evaluation using the predictive mean and variance of the measurements, Z. Left- and right-hand panels have the same meaning as in Figure 3. Four different values for κ_U are shown, given in the panel titles.

in the simulators that we include in the inference. A 'sufficiently large ensemble' allows us to ignore the role of $\mathrm{Var}(\mathcal{M}(X))$ in the inference; it also makes the ensemble variance matrix S a reasonable estimate of $\mathrm{Var}(\mathcal{R}(X))$. (If we want to regularise this estimate then we will need to introduce at least one additional subjective quantity into our inference.) A natural judgement about the relationship between the simulators and the system implies that $\mathrm{Var}(U) > \mathrm{Var}(\mathcal{R}(X)) \approx S$, and this suggests representing $\mathrm{Var}(U)$ as approximately $\kappa_U^2 S$, for $\kappa_U > 1$. Then we are able to compute a mean and variance for the system Y, updated by the values in the ensemble $\mathbb{X}$ and the observations Z. This turns out to be a tractable calculation that can be performed in about a second on an elderly laptop, even for collections of thousands of values, owing to the Bayes linear sufficiency of the ensemble mean.

Where we are slightly less ignorant, we can replace a scalar κ_U with a diagonal matrix, allowing our judgements to vary component-by-component in the collection of values. For example, to vary by latitude, or to be one value for temperature, and another for precipitation, and so on. Where our ensemble is small, or where we have chosen to use only a small subset of simulators, our inference will be sensitive to the choice of $\mathrm{Var}(\mathcal{M}(X))$. The Bayes linear approach we have adopted here can generalise as our judgements become stronger, or when we need to augment our small ensemble with judgements; we have shown only the simplest implementation.

In this analysis, our ignorance may be real or adopted; in the latter case it represents a strategic decision to discard information in the interests of a simpler inference. It may also be political. The IPCC, for example, was set up by the United Nations Environment Programme and the World Meteorological Organisation. As an intergovernmental body with nearly 200 member countries, it is not in the business of treating some climate simulators as *a priori* better than others, once some threshold for adequacy has been passed.

We propose our co-exchangeable model as a general framework for using ensembles of computer simulations to make inferences about complex systems. It represents an innovation in climate science. The current standard approach, which we referred to as the 'Tebaldi model' in section 3, is based on the truth-plus-noise model, which, superficially, appears quite different to our model. We showed in section 2 that in fact the truth-plus-error model is a restricted version of our co-exchangeable model, and we argued there and in section 3 that this restriction represents a strong judgement that has the unfortunate side-effect of preserving hard-to-make subjective assessments, that would have dropped out of the co-exchangeable model.

In complex systems more generally we strongly advocate the use of second-order inference, namely the Bayes linear approach described here. In its simplest form, this approach deals effortlessly with large collections of quantities, and this is also the case for other related inferences (see, e.g. Craig *et al.*, 2001; Goldstein and Rougier, 2004, 2006, 2009; Vernon *et al.*, 2010). Its inferential output, which is expressed in terms of means, variances, and covariances, may seem restrictive, especially when policy interest includes extremes. But it also protects us from the higher-order consequences of implementing fully-probabilistic statistical models that we cannot defend, except in terms of their computational tractability. If probabilities of extremes are required, then they can be bounded by standard inequalities based on means and variances (e.g., the three sigma rule, see Pukelsheim, 1994). When these bounds are large, we should not regard them as unhelpful, but rather as a warning that a tighter bound can only be achieved by higher-moment specifications, about which we may have less well-formed judgements.

References

J.D. Annan and J.C. Hargreaves, 2010. Reliability of the CMIP3 ensemble. *Geophysical Research Letters*, **37**, L02703.

L.M. Berliner and Y. Kim, 2008. Bayesian design and analysis for superensemble-based climate forecasting. *Journal of Climate*, **21**(9), 1891–1910.

P.J. Bickel and E. Levina, 2008. Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**(1), 199–227.

G.E.P. Box, 1980. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**(4), 383–430. With discussion.

C.M. Buser, H.R. Künsh, D. Lüthi, M. Wild, and C. Schär, 2009. Bayesian multumodel projection of climate: bias assumptions and interannual variability. *Climate Dynamics*, **33**, 849–868.

R. Chandler. Exploiting strength, discounting weakness: Combining information from multiple climate simulators. Technical report, Department of Statistical Science, University College London, 2010. Research Report 311.

P.S. Craig, M. Goldstein, J.C. Rougier, and A.H. Seheult, 2001. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, **96**, 717–729.

B. de Finetti, 1937. la prévision, ses lois logiques, ses sources subjectives. *Annals de L'Institute Henri Poincaré*, **7**, 1–68. see de Finetti (1964).

B. de Finetti, 1964. Foresight, its logical laws, its subjective sources. In H. Kyburg and H. Smokler, editors, *Studies in Subjective Probability*, pages 93–158. New York: Wiley. English translation: H. Kyburg.

R. Furrer, S.R. Sain, D. Nychka, and G.A. Meehl, 2007. Multivariate Bayesian analysis of Atmosphere-Ocean General Circulation Models. *Environmental and Ecological Statistics*, **14**, 249–266.

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

T. Gneiting, 1999. Correlation functions for atmospheric data analysis. *Quarterly Joural of the Royal Meteorological Society*, **125**, 2449–2464.

M. Goldstein, 1986. Exchangeable belief structures. *Journal of the American Statistical Association*, **81**, 971–976.

M. Goldstein and J.C. Rougier, 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, **26**(2), 467–487.

M. Goldstein and J.C. Rougier, 2006. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, **101**, 1132–1143.

M. Goldstein and J.C. Rougier, 2009. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, **139**, 1221–1239. With discussion.

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. Chichester, England: John Wiley & Sons.

M. Jun, R. Knutti, and D. Nychka, 2008a. Local eigenvalue analysis of CMIP3 climate model errors. *Tellus*, **60A**, 992–1000.

M. Jun, R. Knutti, and D. Nychka, 2008b. Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *Journal of the American Statistical Association*, **103**, 934–947.

N.R. Kaye, A. Hartley, and D. Hemming, 2011. Mapping the climate: Guidance on appropriate techniques to map climate variables and their uncertainty. *Geoscientific Model Development Discussions*, **4**, 1875–1906. DOI:10.5194/gmdd-4-1875-2011.

R. Knutti, G. Abramowitz, M. Collins, V. Eyring, P.J. Gleckler, B. Hewitson, and L. Mearns. Good practice guidance paper on assessing and combining multi model climate projections. In T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley, editors, *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, 2010a.

R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G.A. Meehl, 2010b. Challenges in combining projections from multiple climate models. *Journal of Climate*, **23**, 2739–2758.

N.A. Leith and R.E. Chandler, 2010. A framework for interpreting climate model outputs. *Applied Statistics*, **59**(2), 279–96.

M.E. McIntyre, 1997. Lucidity and science I: Writing skills and the pattern perception hypothesis. *Interdisciplinary Science Reviews*, **22**(3), 199–216.

G.A. Meehl, T.F. Stocker, W.D. Collins, P. Friedlingstein, A.T. Gaye, J.M. Gregory, A. Kitoh, R. Knutti, J.M. Murphy, A. Noda, S.C.B. Raper, I.G. Watterson, A.J. Weaver, and Z.-C. Zhao, 2007. Global climate projections. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, editors, *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 747–846. Cambridge University Press, Cambridge, UK.

N. Oreskes, 2000. Why predict? Historical perspectives on prediction in Earth Science. In D. Sarewitz, R.A. Pielke Jr, and R. Byerly, editors, *Prediction: Science, Decision Making, and the Future of Nature*, pages 23–40. Island Press.

O.H. Pilkey and L. Pilkey-Jarvis, 2007. *Useless Arithmetic: Why Environmental Scientists Can't Predict the Future*. Columbia University Press.

F. Pukelsheim, 1994. The three sigma rule. *The American Statistician*, **48**, 88–91.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3, `http://www.R-project.org/`.

C.E. Rasmussen and C.K.I. Williams, 2006. *Gaussian Processes for Machine Learning*. MIT Press. Available online at `http://www.GaussianProcess.org/gpml/`.

J.C. Rougier, 2006. Comment on the paper by Haslett *et al.* *Journal of the Royal Statistical Society, Series A*, **169**(3), 432–433.

B. Sansó, C. Forest, and D. Zantedeschi, 2008. Inferring climate system properties using a computer model. *Bayesian Analysis*, **3**(1), 1–38. With discussion, pp. 39–62.

M.J. Schervish, 1995. *Theory of Statistics*. New York: Springer. Corrected second printing, 1997.

R.W. Sinnott, 1984. Virtues of the Haversine. *Sky and Telescope*, **68**(2), 159. Reference from `http://www.usenet-replayer.com/faq/comp.infosystems.gis.html`.

R.L. Smith, C. Tebaldi, D. Nychka, and L.O. Mearns, 2009. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, **104**(485), 97–116.

O. Talagrand, 1997. Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan*, **75**(1B), 191–209.

C. Tebaldi and R. Knutti, 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2053–2075.

C. Tebaldi and B. Sansó, 2009. Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *Journal of the Royal Statistical Society, Series A*, **172**(1), 830–106.

C. Tebaldi, R.L. Smith, D. Nychka, and L.O. Mearns, 2005. Quantifying uncertainty in projections of regional climate: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, **18**, 1524–1540.

C. Tebaldi, R.L. Smith, and B. Sansó. Characterizing uncertainty of future climate change projections using hierarchical Bayesian models. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P Dawid, D. Heckerman, A.F.M Smith, and M. West, editors, *Bayesian Statistics 9*. Oxford: Oxford University Press, 2010.

I. Vernon, M. Goldstein, and R.G. Bower, 2010. Galaxy formation: a Bayesian uncertainty analysis. *Bayesian Analysis*, **5**(4), 619–670.