Bayesian Analysis (2015) **10**, Number 4, pp. 877–908

Posterior Belief Assessment: Extracting Meaningful Subjective Judgements from Bayesian Analyses with Complex Statistical Models

Daniel Williamson* and Michael Goldstein†

Abstract. In this paper, we are concerned with attributing meaning to the results of a Bayesian analysis for a problem which is sufficiently complex that we are unable to assert a precise correspondence between the expert probabilistic judgements of the analyst and the particular forms chosen for the prior specification and the likelihood for the analysis. In order to do this, we propose performing a finite collection of additional Bayesian analyses under alternative collections of prior and likelihood modelling judgements that we may also view as representative of our prior knowledge and the problem structure, and use these to compute posterior belief assessments for key quantities of interest. We show that these assessments are closer to our true underlying beliefs than the original Bayesian analysis and use the temporal sure preference principle to establish a probabilistic relationship between our true posterior judgements, our posterior belief assessment and our original Bayesian analysis to make this precise. We exploit second order exchangeability in order to generalise our approach to situations where there are infinitely many alternative Bayesian analyses we might consider as informative for our true judgements so that the method remains tractable even in these cases. We argue that posterior belief assessment is a tractable and powerful alternative to robust Bayesian analysis. We describe a methodology for computing posterior belief assessments in even the most complex of statistical models and illustrate with an example of calibrating an expensive ocean model in order to quantify uncertainty about global mean temperature in the real ocean.

Keywords: prevision, subjective Bayes, temporal sure preference, Bayesian analysis, MCMC.

1 Introduction

The idea that uncertainty is a subjective property of individuals underlies the development of the field of Bayesian statistics (cf. Savage, 1977; Lindley, 2000). You are uncertain about some aspect of the world, you specify this uncertainty in the form of a prior probability distribution. You specify a probability model describing the data generating process that you will use to learn about the world and update your prior

*Statistical Science College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, d.williamson@exeter.ac.uk, url: https://emps.exeter.ac.uk/mathematics/staff/dw356

†Department of Mathematical Sciences Durham University, Durham, UK, michael.goldstein@durham.ac.uk

 DOI: 10.1214/15-BA966SI

probabilities to posterior probabilities, upon observation of the data, using Bayes theorem. We term this the subjective Bayesian approach.

Many aspects of a Bayesian analysis are challenging, in particular, when a prior and likelihood have been established, computing or sampling from the posterior distribution can be extremely difficult. This challenge, however, has been well met by the Bayesian field, with modern techniques in Markov Chain Monte Carlo (MCMC) allowing even the most complicated statistical models to be sampled from in order to obtain posterior probabilities (see, for example, Liang et al., 2010). However, given the complexity of the statistical models that we are able to develop, the demands of the science that they apply to, and the power of our computational methods for sampling from them, methodology for the elicitation of subjective prior probabilities lags behind and remains extremely difficult and highly controversial.

For these reasons and, perhaps, others, the field of so called "objective" Bayesian analysis (or O-Bayes), is now extremely popular and represents, arguably, the most popular approach to Bayesian inference. O-Bayes seeks to develop and use automatic prior distributions for any probability model that, in some sense depending on the type of prior developed, have a minimal influence on the output of a Bayesian analysis. For details of the approach see Berger (2006), for a discussion of the competing philosophies between O-Bayes and subjective Bayes see Goldstein (2006); Berger (2006) and the following discussion. This paper concerns the development of the subjective Bayesian approach, hence we do not comment further on O-Bayes solutions or philosophies in the rest of our narrative.

The statistical models we are now able to develop would, if we were to attempt a fully subjective Bayesian analysis, often require elicitation of high dimensional joint prior distributions, perhaps over spatio-temporal fields or over parameters of multiple types, and perhaps involving non-standard distributional forms. Current state-of-the-art elicitation frameworks and tools such as SHELF (Oakley and O'Hagan, 2010) and MATCH (Morris et al., 2014), enable the elicitation of univariate distributions of standard forms such as Normal, Beta and Gamma distributions and do not come close to meeting the requirements made for some statistical models. Even univariate parameters in a complex hierarchical statistical model can be subject to so many layers of abstraction that even understanding what they mean or how they might impact upon an analysis can make elicitation using current tools challenging or even dubious. Even if elicitation methodology were substantially more advanced, time and budget constraints in many investigations may prohibit the sort of elicitation that involves either the statistician, the experts or both believing every probability statement made in order to facilitate the Bayesian analysis.

Our argument then is that even if elicitation methodology catches up with the computational capabilities of the field and the demand of the problems we study, in most cases we will not hold any prior probability distribution or even the likelihood as representative of our beliefs in complex problems. What meaning can we then give to any Bayesian analysis that we do perform? Most importantly, is there anything we can conclude about our own uncertainty from a Bayesian analysis?

Bayes linear methods (Goldstein and Wooff, 2007) are an offered solution to this problem. The idea is based on making only partial subjective prior specifications in the form of means, variances and covariances, and using the geometry of Hilbert spaces to update these beliefs using data. Whilst attractive and computationally tractable in many applications, these methods do not allow us to take advantage of the many advanced technologies and benefits of the modern fully probabilistic Bayesian approach, nor do they allow us to combine and update any fully probabilistic judgements for certain aspects of our statistical model that we might hold. We also comment that the fully Bayesian approach is well established and it is therefore easier to explain to and ultimately publish with collaborators from other fields.

Robust Bayes was a potential avenue of study into this problem. Popular in the 1980s and 1990s, robust Bayes looked to explore whether any conclusions from a Bayesian analysis were robust to classes of prior and likelihood choices (see Berger, 1994, for a nice overview). However, the method focussed on analytic solutions in relatively simple Bayesian models and, as the computational capabilities of the field increased exponentially and Bayesian models became more complex, robust Bayesian approaches became intractable and are now rarely pursued.

Robust Bayes can be thought of as a formalism of sensitivity analysis, whereby alternative Bayesian analyses are considered to explore how sensitive posterior inferences are to prior modelling choices (see, for example, Gelman et al., 2004, chapter 5). Though sensitivity analysis is an important and useful step in any Bayesian analysis, there is no formal mechanism for arriving at posterior judgement following it. For example, what do samples from any alternative posterior distributions obtained during the sensitivity analysis say about your actual judgements? Do these alternative posteriors, or indeed that which formed the main analysis, represent subjective probabilities? In addition to this, there is no formal method for ensuring a sensitivity analysis is "complete" in the sense that it fully explores all possible alternative analyses that might have been deemed reasonable by the analyst and expert.

In this paper, we describe a new methodology, which we term *posterior belief assessment*, that aims to improve a full Bayes analysis that we have performed (or can perform) with prior distributions and likelihood (judgements) that represent our best current judgement (without our necessarily believing every probability statement made by these judgements). By best current judgement here, we mean that we have expert probabilistic judgements that we are unable to adequately express and that the chosen combination of prior and likelihood represent these in some way but, due to the complexity of the problem, we are unable to assert a precise correspondence between the chosen judgements and our actual beliefs.

Our methodology improves our Bayesian analysis through performing further calculations under alternative judgements, in order to get closer to our actual posterior beliefs for key quantities of interest in a measurable way. Similar to robust Bayes, we attempt to consider all possible alternative forms that we might give to these judgements. However, our approach is to use this information along with foundational arguments, to derive our subjective judgements for key quantities rather than to look for any posterior probability statements that are robust to these choices.

Before proceeding, we comment that Bayesian model averaging (BMA, Hoeting et al., 1999), makes similar arguments regarding the existence of alternative judgements (or models). That approach, which puts a probability distribution over a class of alternative models, is, however, fundamentally different to what is described here. By putting a subjective probability distribution over any alternative judgements we consider, as in BMA, we would, under a subjectivist definition of probability, be making the extremely strong statement that there is a true model within the set considered. We already know that this is false in problems of reasonable complexity, such as that given above. Our approach will not make assumptions of this type. Instead, we begin with the foundations of subjective inference to establish a relationship between our underlying judgements and our Bayesian analysis (Section 3). We then use this relationship to develop a methodology for combining an original analysis with alternatives to get closer to those judgements (Sections 4 and 5).

We begin, in Section 2, by setting up a motivating example involving the Bayesian calibration of an expensive ocean model for learning about the real ocean temperatures. In Section 3, we discuss the meaning of a subjective Bayesian analysis and develop the foundational arguments and machinery needed in order to develop our methodology. Section 4 describes our approach in the case of finite, quantifiable, alternative modelling judgements. Section 5 extends the approach to the case where we have infinitely many alternative modelling judgements by using co-exchangeability to partition the space of alternatives. Section 6 illustrates the application of posterior belief assessment to the example introduced in Section 2, and Section 7 contains discussion.

2 A motivating example

Throughout the paper we will discuss current judgements and alternative judgements regarding our models and specification for the prior and likelihood in complex statistical problems, and we will claim that perhaps none of these actually reflect our internal expert probabilistic judgements on the observables or any key elements of the problem for which we intend to make probabilistic inference (if such probabilistic judgements could ever exist or be obtained if they did). To motivate this discussion, we introduce an example from our own work that is sufficiently complex to illustrate these ideas and to make concrete some of our terminology. We apply our methodology to this example in Section 6.

Our example involves the calibration of a computationally expensive ocean model used to learn about the state of the real ocean. We describe the particular ocean model we are using and our experiments on it in detail in Section 6. To keep things more general initially, we will denote the ocean model $f(x, d)$ with x representing the model input parameters chosen for a particular run and $f(\cdot, d)$ representing the model output of interest, which, in this case will be the global mean temperature at depth d (in fact, the model will output 6 hourly values for temperature, velocity and salinity over a 3D mesh covering the globe at 31 depth levels, and we post-process it to only work with outputs of interest).

Computer model calibration (developed in Kennedy and O'Hagan, 2001) is a method

for combining computer model output and real world data in order to probabilistically describe past, current or future states of the real world. To motivate our methodology, we describe a statistical model for calibrating our ocean model using observed temperature at a number of depths in order to predict temperature at an unobserved depth. What follows is a brief outline of the statistical modelling involved in calibration in order to set up a Bayesian network (BN) that enables us to easily talk about required judgements and not ruled out alternatives, and to have an example in mind when we develop the underlying theory behind our approach. We do not reproduce the details of the calculations required to obtain conditionals for posterior sampling via MCMC. These are available in Kennedy and O'Hagan (2001).

The modelling begins with observations $z(d)$ at depth d of components of the ocean y made with independent error e so that

$$z(d) = y(d) + e; \qquad e \sim \mathrm{N}(0, \sigma_e^2).$$

These elements relating to the real world only appear in blue on our BN shown in Figure 1.

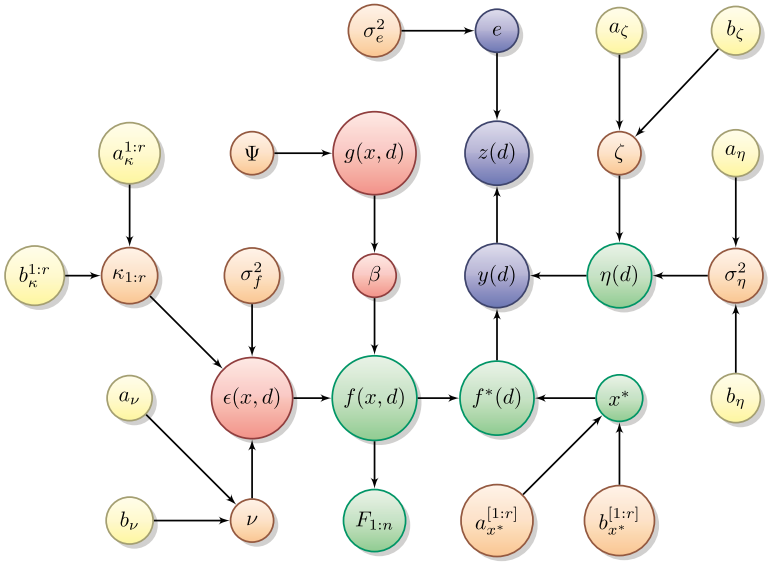

Figure 1: Bayesian network for the ocean model calibration problem.

Components relating the ocean model to the ocean appear in green in Figure 1. The model refers to the best input approach, which indicates that there is a particular setting of the inputs, x^*, that is sufficient for the ocean model in informing us about reality at any depth d. The model is

$$y(d) = f(x^*, d) + \eta(d);$$

with discrepancy $\eta(d)$ independent from x^* and $f(x,d)$ for all x. The model runs that inform us about $f(x,d)$ are denoted $F_{1:n}$. Next we use a statistical model called an emulator to model our uncertainty about $f(x,d)$. Components of the emulator appear as the red nodes in Figure 1. The emulator is

$$f(x,d) = \beta^T g(x,d) + \epsilon(x,d); \qquad \epsilon(x,d) \sim GP(0, c_f(\cdot,\cdot))$$

with $g(x,d)$ a chosen vector of basis functions, β uncertain coefficients and $\epsilon(x,d)$ an independent (from β) mean zero Gaussian process with specified covariance function $c_f((x,d),(x',d'))$. We usually take this to be weakly stationary so that $c_f((x,d),(x',d')) = c_f(|(x,d)-(x',d')|)$.

Parameters that must be specified to characterise our uncertainty for each model component described so far appear on the BN as orange nodes. Starting with the discrepancy, $\eta(d)$, Kennedy and O'Hagan (2001) suggest a mean zero weakly stationary Gaussian process prior with Gaussian covariance function $c_\eta(d,d') = \sigma_\eta^2 \exp\{-\zeta|d-d'|^2\}$, so that a variance parameter σ_η^2 and a depth correlation parameter ζ are introduced. The best input x^* is given a uniform distribution for each of the r dimensions of the input space with lower and upper bounds, $a_{x^*}^{[i]}$ and $b_{x^*}^{[i]}$.

We will build separate univariate emulators for each depth d, altering the specification from one of requiring $\beta^T g(x,d)$ and $\epsilon(x,d)$ to one requiring $\beta_d^T g_d(x)$ and $\epsilon_d(x)$. For the response surface $\beta_d^T g_d(x)$ we restrict the prior specification to one of choosing the number and type of basis functions to enter into each $g_d(x)$, choices we identify with the symbol Ψ and whose options will be detailed in Section 6.2. We specify a joint prior $p(\beta, \sigma_f^2)$ as described in Section 6.2. For the Gaussian process residual, $\epsilon_d(x)$, we let

$$c(x,x') = \sigma_f^2 \left(\nu \mathbb{1}_{x=x'} + (1-\nu) R(|x-x'|; \kappa_{1:r})\right)$$

where ν, traditionally called the "nugget" (Andrianakis and Challenor, 2012), represents the proportion of the residual variability that is, in this case, due to internal variability of the climate model (see Williamson and Blaker, 2014, for discussion) and $R(|x-x'|;\kappa_{1:r})$ is a correlation function depending on roughness parameters κ on each dimension of the input space. We say more on the possible choices of correlation function below.

For each of the model parameters, we might be uncomfortable in expressing values directly and could put hyper-priors on each. In our Bayesian network we do this for 4 of the more difficult parameters to consider: the correlation lengths and variance of the discrepancy, ζ and σ_η^2, the nugget term ν and the r correlation parameters, $\kappa_{1:r}$, of the emulator residual. We choose

$$\zeta \sim \mathrm{G}(a_\zeta, b_\zeta), \qquad \sigma_\eta^2 \sim \mathrm{IG}(a_\eta, b_\eta), \qquad \nu \sim \mathrm{Be}(a_\nu, b_\nu),$$

and use separate $\mathrm{Be}(a_\kappa^i, b_\kappa^i)$ priors for each half length correlation of the emulator residual. A half length correlation for the ith input between the correlated parts of $\epsilon(x)$ and $\epsilon(x')$ is the value of $R(|x-x'|)$ when all elements of x and x' are equal with the exception of x_i and x_i' and where $|x_i - x_i'|$ is equal to half of the possible range of x_i (see Higdon et al., 2008; Williamson and Blaker, 2014; Williamson, 2014, for further details). Hyper-parameters appear on our BN in yellow.

The Bayesian network for this problem in Figure 1 required many complex probabilistic modelling judgements just to write down, and requires a great deal more specification before a Bayesian analysis can be performed. Values for all hyper-parameters and other parent nodes are required to specify prior distributions, and modelling choices such as the form of the correlation function for the Gaussian process are required before we have established a likelihood. Even at this stage, the required modelling and specification is, perhaps, too advanced to expect an ocean modeller, or even a statistician-modeller team to ever be able to complete the specification in such a way so that every probabilistic statement we have made about any collection of nodes on the graph represents what they really believe.

In fact, as we believe would often be the case with a complex statistical model, it is very easy to criticise the modelling we have done as not representative of the beliefs of any expert ocean modeller. We have stated already that in order to complete a subjective elicitation in this problem directly from the expert, we would require an ocean modeller with a deep understanding of probabilistic reasoning and the implications of this type of statistical model. However, even if such an expert did exist and had enough time to spend with us, there are a number of criticisms they could level at the model right off the bat. First, there is the idea that the discrepancy is additive and independent of the model. This is problematic because if we knew x^*, our ocean reconstruction y would not be dynamically consistent, i.e. it would not obey the Navier–Stokes equations over a reasonably short time period. This is because our modelling of η does not require that it does (nor could it ever, reasonably). We might visit every part of our modelling and point out holes in the realism, though we satisfy ourselves with one further example. We know that the residual from the response surface we fit using our emulator is not really a realisation of a weakly stationary Gaussian process. Though, if we are lucky, we might find a reasonable fit using such a process on a scale similar to the average distance between our design points, it is unlikely that this model adequately captures the behaviour of the function at, for example, much smaller scales in all parts of the parameter space. The assumption of stationarity here is one made largely for computational convenience and one made because experience in this type of modelling has often shown that breaking this assumption does not gain a great deal in terms of predictive power relative to the effort required in its implementation.

There are many judgements already made that we might view as equivalent to or "just as good as" alternative judgements that we might have made (for example, we could have chosen a lognormal prior for ζ instead of a gamma). There are many more we are yet to make to which the same argument will apply and, we know, for any alternative set of judgements, that the framework we are using is only a model for our actual underlying judgement. We make this precise in our theoretical development below.

3 Interpreting a Bayesian analysis

There are different views as to the primitive concepts underpinning the Bayesian approach to inference. In this paper, we are driven by the requirement to give a formal

and operational treatment which deals with the difference between the inferences made within a traditional Bayesian model for a complex problem and the actual inferences that we are confident to assert on the basis of our modelling and analysis. It is easy to see why there will often be a mismatch between these two levels of inference. The full Bayes model requires prior specification over complex high dimensional spaces for which we may struggle to make meaningful judgements to the extreme level of detail required for the analysis. Lacking confidence in our prior assessments and in the many, often somewhat arbitrary, modelling choices that we have made in order to complete our analysis, it is hard to have confidence in our posterior judgments. We need to express our uncertainty about the relevance of our Bayesian modelling to our actual real world inferences. However, an attempt to make such a higher level uncertainty specification through the medium of probability is, in many ways, more difficult than was the original modelling task itself. Partly, this difficulty is intrinsic to the task itself. However, partly this is because probability, due to the exhaustive level of detail that is required in the analysis, is not an appropriate primitive, in our view, to bridge the gap between a formal Bayes analysis and the actual posterior judgements that we may wish to make.

Therefore, in this paper, we treat expectation, rather than probability, as the primitive for this higher level analysis. In this way, we may still consider any probabilities that are of direct concern, as probabilities, in this view, are simply the expectation of the corresponding indicator functions. However, we have the option of restricting the level of detail of our specification, by just considering those expectation specifications which are directly relevant to the problem at hand. We shall show, with expectation as primitive, how to develop a formal and operational approach to addressing the gap between Bayesian inference and actual real world inference for complex models, and illustrate the practicality of the approach in application to a model of realistic size and complexity.

It would be interesting to compare our approach with a corresponding alternative developed purely within the formalism of probability, but we do not know of any practical way to do this. If we were to choose probability as primitive, then we would need to assess a full probabilistic specification describing the relationship between the Bayes analysis and our judgements, the specification of which would be even more difficult to attribute precise meaning to than the probabilistic specification underpinning the original Bayes analysis.

3.1 Expectation as primitive

In the fundamental volumes summarising his life's work (de Finetti, 1974, 1975), de Finetti based the theory of probability on expectation, or, as he termed it, *prevision*, as the primitive for the theory. de Finetti gives the following operational definition for your expectation, $\mathrm{E}[X]$, for the random quantity X, namely that it is your preferred choice for the value of c when confronted with the penalty $(X-c)^2$, where the pay-off is in some appropriate units, for example, probability currency. (In this theory, all expectations are the subjective judgements of specified individuals. There is no place for the notion of objective probability, beyond, informally, the common language meaning of a consensus judgement shared among a group of individuals.) Of course, in practice, we may choose

many alternative ways to elicit your choice for $\mathrm{E}[X]$, but this definition is particularly appropriate when deriving the formal properties of expectation as it leads directly to the geometric structure underpinning expectation based approaches to inference (see, for example, Goldstein and Wooff, 2007, chapter 3).

Within this formulation of probability theory, we may derive an account of the relationship between conditional expectation, $\mathrm{E}[X|D]$ given observation of a member of the partition $D = (D_1, \ldots, D_k)$, for which one and only one of the outcomes D_i will be observed, and the actual expectation, or prevision, $P_t(X)$ that we may specify at the future time t when we do observe the outcome of the partition (We use $P_t(X)$ rather than $\mathrm{E}_t[X]$ here, as the latter is reserved for an operation that we will define later and make a great deal of use of. Only the currently unobtainable time t prevision is denoted with P. For all other subjective expectations we use the more conventional E operator). This development was first described in Goldstein (1997) and the remainder of this section gives a summary of it.

As the conditional probability of an event is equivalent to the conditional expectation for the indicator function for the event, this also gives an account of the relationship between conditional probabilities, for example, assessed by Bayes theorem, and actual posterior probabilities when the conditioning event is observed. It is important to maintain the distinction between conditional probabilities, defined in subjective probability theory as bets which are called off unless the conditioning event occurs, and posterior probabilities which you assign when you have seen the corresponding event. There is no obvious formal relationship between these two notions whatsoever, and foundational descriptions typically restrict the formal development to perfectly rational individuals, operating in small and very tightly constrained worlds. As such, conditional probability may be interpreted as a simple and flexible model for real world inference. However, like any model, it is essential to be careful in considering the relationship between the model and the real world, which is not tightly constrained and concerns the actual inferences of real individuals.

Although there is no deterministic relationship between conditional and posterior probabilities, we can derive certain probabilistic relationships between the two concepts. Specifically, we may use the notion of temporal sure preference. Suppose that we must choose between two random penalties A and B. We say that we have a sure preference for A over B at (future) time t if we are sure now that, at time t, we will prefer A to B. The temporal sure preference (TSP) principle says that if we have a sure preference for A over B at t, then we should not have a preference for B over A now. Temporal sure preference is a very weak property (as it is hard for us to hold temporal sure preferences). However, it is sufficient to derive the basic probabilistic relationship between conditional and posterior expectations, which is as follows.

Suppose that we currently specify our conditional expectation for a random vector X given partition $D = (D_1, \ldots, D_k)$. Suppose, at future time t when we have observed which element of D occurs, we make an actual posterior expectation statement $P_t(X)$. If we denote by $\mathrm{E}[X|D]$ the random quantity which takes value $\mathrm{E}[X|D_i]$ if we observe outcome D_i, then given TSP we can show (see Goldstein, 1997, for details) that, now,

we must have the following orthogonal decomposition:

$$X = (X - P_t(X)) \oplus (P_t(X) - \mathrm{E}[X|D]) \oplus (\mathrm{E}[X|D] - \mathrm{E}[X]) \oplus \mathrm{E}[X]. \quad (1)$$

Each of the three bracketed terms on the right hand side of (1) has expectation zero, and the three terms are mutually uncorrelated (the notation $A \oplus B$ means that the random vectors A and B are uncorrelated). More strongly, $(P_t(X) - \mathrm{E}[X|D])$ has expectation zero conditional on each member of the partition. We may view $\mathrm{E}[X|D]$ as a prior inference for $P_t(X)$, in the sense that $\mathrm{E}[X|D]$ bears the same relationship to $P_t(X)$ that $P_t(X)$ bears to X. In this way, we can make precise the notion that the formal Bayes analysis is a model for your posterior judgements, whose relationship is the same as for any model of a real world process (Goldstein, 2011).

As the terms in (1) are uncorrelated, we can interpret this construction in terms of the resulting variance partition,

$$\mathrm{Var}[X] = \mathrm{Var}[X - P_t(X)] + \mathrm{Var}[P_t(X) - \mathrm{E}[X|D]] + \mathrm{Var}[\mathrm{E}[X|D]]. \quad (2)$$

The variance of X is decomposed into three components. $\mathrm{Var}[\mathrm{E}[X|D]]$ expresses the information content of the formal Bayes analysis, $\mathrm{Var}[P_t(X) - \mathrm{E}[X|D]]$ expresses the information content of the further judgements and information, in addition to the formal Bayes analysis, that we may bring to bear by time t, and $\mathrm{Var}[X - P_t(X)]$ expresses the intrinsic limitations to our inferences by this time. We can expand the value of the Bayesian component of the analysis by expanding the partition and thus the variation attributable to the conditioning.

While we have described the variance partition for a full Bayes analysis, the argument from TSP actually establishes the more general property that for any random quantity W which we will certainly observe by time t, then we must assign

$$\mathrm{E}\left[(X_i - P_t(X_i))^2\right] \leq \mathrm{E}\left[(X_i - W)^2\right].$$

By varying the choice of W, we can derive the corresponding representation

$$X = (X - P_t(X)) \oplus (P_t(X) - \mathrm{E}_U[X]) \oplus (\mathrm{E}_U[X] - \mathrm{E}[X]) \oplus \mathrm{E}[X] \quad (3)$$

where $\mathrm{E}_U[X]$ is the orthogonal projection of X into the collection of linear combinations of the elements of the collection $U = (W_1, \ldots, W_k)$ where each W_i is a random quantity which will be observed by time t, and the terms on the right hand side of (3) obey the same orthogonality properties as those of (1), with a similar interpretation.

In this next section, we shall combine these two variance partitions to give a general representation of the inference that we may make through a formal Bayes analysis for a complex problem. Firstly, (1) is based on the requirement that all of the conditioning statements refer to your actual probabilistic judgements. In practice, for complex problems, the level of detail required is too extreme for this to be possible. We should instead view the choices made in the full Bayes analysis as providing a model for your actual probabilistic judgements (in a way we shall make clear in the next section),

which is itself a model for your posterior inferences. Therefore, we will consider a range of modelling choices for likelihoods, priors and so forth. Each choice leads to a Bayesian conditional expectation. We now use (3) to construct the corresponding relationship between the outcomes of each collection of modelling choices that have been considered and the actual inferences that you are able to make given the data and your collection of analyses.

4 Posterior belief assessment

Define a vector of quantities of interest, y, so that your principal interest in your Bayesian analysis is in finding posterior expectations of y given relevant data z. Note that this definition allows us to be interested in posterior variances and posterior probabilities of key events (defined as expectations of indicator functions), for example, the probability that some quantity exceeds a threshold. Define your current modelling judgements as the set J_0. These judgements comprise everything required to perform the Bayesian calculation, including form and any hyper-parameters of both the prior and the likelihood. We do not believe all of the probability statements in J_0, but we do view J_0 as "representative enough" of the structure of the problem and any scientific judgements and beliefs that we do hold, to allow us to perform the Bayesian computation and to believe that $\mathrm{E}[y|z; J_0]$ is informative for our posterior prevision for y. As discussed in Section 3, this conditional expectation will not be your prevision. Operationally, this means that we prefer penalty $(y_i - P_t(y_i))^2$ to $(y_i - \mathrm{E}[y_i|z; J_0])^2$, equivalently

$$\mathrm{E}\left[(y_i - P_t(y_i))^2\right] \leq \mathrm{E}\left[(y_i - \mathrm{E}[y_i|z; J_0])^2\right].$$

In Section 3, we argued that nothing we can do prior to time t will give us $P_t(y)$ at time t after seeing z. In particular, we cannot set up a called off bet via our Bayesian analysis under J_0 and be required to hold $\mathrm{E}[y|z; J_0]$ at time t, even if we really believed all of the probability statements in J_0. However, though $P_t(y)$ may be unobtainable within the constraints of the Bayesian formalism that we have expressed, we would like to get as close as possible to it using the fully Bayesian machinery. For example, can we find good choices of $\Gamma(z)$ to improve upon $\mathrm{E}[y|z; J_0]$ (in the sense that you prefer penalty $(y_i - \Gamma_i(z))^2$ to $(y_i - \mathrm{E}[y_i|z; J_0])^2$ for any vector-valued function Γ of the data at time t, for all i)?

Though the J_0 contains sufficient structure to enable us to perform a Bayesian analysis that we view as informative for certain posterior expectations over quantities of interest, as we do not believe all of the probability statements made by J_0, we do not view these judgements as uniquely representative of the problem structure and our scientific insight. By changing parts of the model and/or the prior (or even the way we generate posterior samples) in a way that either represents our scientific insights and beliefs in a different way, or that changes them in such a subtle way that we would still view the analysis as informative for our prevision for y, we can arrive at a set of alternative judgements $J_1, \ldots, J_k$. We could, in principle, run alternative Bayesian analyses under each set of judgements and compute $\mathrm{E}[y|z; J_i]$ for $i = 1, \ldots, k$. We

assume a finite collection of alternative judgements we would adopt here, and relax this constraint in Section 5.

We note here that in viewing the Bayesian analysis like this, we are using probability in two different ways. The first, in the true subjective sense, represents what we actually believe about a quantity, and these are our previsions. The second uses probability as a useful modelling language for transforming our judgements into previsions. The probability statements made as part of any J_i are therefore not belief statements, nor do we assume that there is an underlying true representation of our beliefs, J_* say, that could be elicited if we could think hard enough for long enough to extract it. However, it is reasonable to talk of potential sets of judgements not representing our beliefs well enough, so that we could rule them out as being informative for our previsions. For this reason, we will sometimes refer to the set of alternative judgements $J_1, \ldots, J_k$ as not ruled out judgements.

That we have chosen J_0 over any $J_1, \ldots, J_k$, may lead us to the view that we prefer penalty $(y_i - \mathrm{E}[y_i|z; J_0])^2$ to $(y_i - \mathrm{E}[y_i|z; J_i])^2$ for $i = 1, \ldots, k$. However, some of the elements of J_0 may have been chosen for pragmatic reasons. For example, we might choose a conjugate prior family to reduce the computational burden in part of our posterior sampling scheme. Even if we were to adopt the view that J_0 leads to conditional expectations that we do prefer to conditional expectations derived through Bayesian analyses under any of the alternative sets of judgements, there may be derivable random penalties that make use of the collection of our current and alternative judgements that we prefer over any based on one set of them. We now consider combinations of quantities related to our alternative judgements, $J_1, \ldots, J_k$. We show that there is a linear combination of the collection of conditional expectations that is at least as close to your actual prevision as $\mathrm{E}[y|z; J_0]$ and derive its properties. Define $\mathcal{G}$ to be the vector $(\mathrm{E}[y|z; J_0], \mathrm{E}[y|z; J_1], \ldots, \mathrm{E}[y|z; J_k]) = (\mathcal{G}_1, \ldots, \mathcal{G}_{k+1})$. Let $\mathcal{G}_0$ be the unit constant. We have the following result.

Theorem 1. *Let*

$$\mathrm{E}_{\mathcal{G}}[y] = \mathrm{E}[y] + \mathrm{Cov}[y, \mathcal{G}] \,\mathrm{Var}[\mathcal{G}]^{-1} (\mathcal{G} - \mathrm{E}[\mathcal{G}]), \tag{4}$$

an expectation which we term, our posterior belief assessment for y. Then

(i) $\mathrm{E}_{\mathcal{G}}[y]$ *is at least as close to y as* $\mathrm{E}[y|z; J_0]$. *Equivalently, for each i,*

$$\mathrm{E}\left[(y_i - \mathrm{E}_{\mathcal{G}}[y_i])^2\right] \leq \mathrm{E}\left[(y_i - \mathrm{E}[y_i|z; J_0])^2\right].$$

where $\mathrm{E}_{\mathcal{G}}[y_i]$ *is the ith component of* $\mathrm{E}_{\mathcal{G}}[y]$.

(ii) $\mathrm{E}_{\mathcal{G}}[y]$ *is at least as close to* $P_t(y)$ *as* $\mathrm{E}[y|z; J_0]$. *Equivalently, for each i,*

$$\mathrm{E}\left[(P_t(y_i) - \mathrm{E}_{\mathcal{G}}[y_i])^2\right] \leq \mathrm{E}\left[(P_t(y_i) - \mathrm{E}[y_i|z; J_0])^2\right].$$

Proof. Our posterior belief assessment is the Bayes linear rule for y, given $\mathcal{G}$ (see Goldstein and Wooff, 2007). For each i, $\mathrm{E}_{\mathcal{G}}[y_i]$ is the Bayes linear rule for y_i, namely the

linear combination $\sum_{i=0} \alpha_i \mathcal{G}_i$ that minimises

$$\mathrm{E}\left[\left(y_i - \sum_{i=0} \alpha_i \mathcal{G}_i\right)^2\right].$$

(see page 56 in Goldstein and Wooff, 2007, for a proof of this result). Therefore, property (i) follows immediately.

From TSP, $y - P_t(y)$ has mean zero and is uncorrelated with all random quantities that will be known by time t, and, in particular, all the elements of $\mathcal{G}$. Therefore, for each i,

$$\begin{aligned} \mathrm{E}\left[(y_i - \mathrm{E}_{\mathcal{G}}[y_i])^2\right] &\leq \mathrm{E}\left[(y_i - \mathrm{E}[y_i|z; J_0])^2\right] \quad \Rightarrow \\ \mathrm{E}\left[(y_i - P_t(y_i))^2\right] + \mathrm{E}\left[(P_t(y_i) - \mathrm{E}_{\mathcal{G}}[y_i])^2\right] &\leq \mathrm{E}\left[(y_i - P_t(y_i))^2\right] \\ &\quad + \mathrm{E}\left[(P_t(y_i) - \mathrm{E}[y_i|z; J_0])^2\right], \end{aligned}$$

and property (2) follows immediately. □

Our theorem identifies the key quantities that we need to assess in order to implement the method, namely $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$. In Section 4.2, we present a method for evaluating these quantities. Note that the theorem implies that by computing the posterior belief assessment using additional experiments, we resolve an additional proportion of our uncertainty in y, compared with the case where we only have one Bayesian analysis. We can estimate a lower bound on the proportion of uncertainty reduced by observing the ratio of the adjusted variances $\mathrm{Var}[y_i - \mathrm{E}_{\mathcal{G}_1}[y_i]]$ and $\mathrm{Var}[y_i - \mathrm{E}_{\mathcal{G}}[y_i]]$ (note $\mathrm{E}_{\mathcal{G}_1}[y_i]$ is computed using (4) replacing $\mathcal{G}$ with $\mathcal{G}_1$). This is a direct corollary of part (i).

Readers familiar with Bayesian Model Averaging (BMA, Hoeting et al., 1999), may notice that (4), which is simply a constant plus a linear combination of each alternative posterior expectation, may be interpreted as a constant plus the BMA posterior expectation under a posterior distribution over model J_i equal to W_{i+1} with

$$W_j = (\mathrm{Cov}[y, \mathcal{G}]\,\mathrm{Var}[\mathcal{G}]^{-1})_j.$$

However, our approach is not the same as we have made expectation primitive and not probability. In BMA, one needs a probability distribution over all possible models in order to proceed. Hence, in this context, the BMA cure to needing to make so many judgements in a full Bayesian analysis and to not being sure that any of these judgements represent our actual beliefs, is to make infinitely more judgements (in the form of a probability distribution across all possible models and conditional probabilities under each model), and proceed from there. But now we are back in the same boat in which we started, requiring many more prior judgements in order to do the Bayesian analysis. We can ask the same questions regarding whether we actually hold each of these judgements, whether the statistical modelling is representative of them, or if there are alternatives that we feel should be explored and so on. More subtly, as mentioned in the introduction, BMA must also assume that one of our alternative models is the truth, and we are unwilling to make this assumption.

With expectation as primitive in a posterior belief assessment, we avoid the infinite regress. We make no probabilistic prior statements about any of our alternative models, hence do not assume any represent our true underlying beliefs, and simply seek to get closer to actual prevision, $P_t(y)$, than $\mathrm{E}[y|z; J_0]$ via orthogonal projection of y onto a vector $\mathcal{G}$ of alternative Bayesian analyses. Importantly we have established the relationship between our posterior belief assessment, $\mathrm{E}_{\mathcal{G}}[y]$, and what we will actually think at time t, $P_t(y)$, which would be a task still ahead of an analyst adopting BMA.

4.1 Practical considerations

The above account states that if we can write down a finite set of alternative judgements $J_1, \ldots, J_k$, to J_0, and if we can obtain conditional expectations for key quantities, y, given data z, using a full Bayesian analysis under each J_i, $i = 0, \ldots, k$, then we can perform a posterior belief assessment to get closer to our actual judgements if we can compute (4). However, there are a number of practical considerations to doing this, even if we can write down $J_1, \ldots, J_k$.

One such consideration is that each alternative Bayesian analysis must be performed in order to obtain each $\mathrm{E}[y|z; J_i], i = 1, \ldots, k$, so that we may evaluate $\mathcal{G}$. There are, essentially, two factors impacting upon the feasibility of this step. The first involves the availability of extra computer power. For many of the J_i, we may only have small changes to hyper-prior quantities to make, or alternative, yet trivially computable, formulations of the likelihood. In these cases, performing the alternative Bayesian computations (e.g. via MCMC) involves little or no manpower and instead requires more computing. In these cases the alternative analyses can run in parallel either on single multi-core machines, or via clusters or distributed computing, representing little additional complexity to the overall analysis. The second factor to consider regards the additional complexity of the Bayesian computation for any of the alternative judgements J_i. An alternative model may lead to a Bayesian calculation that we don't yet know how to do well, or that would take so much additional effort to implement that we regard performing this alternative analysis in order to provide a posterior belief assessment to be either impractical (we do not judge it to be a good use of our time), or infeasible given our current budget constraints. We visit a solution to this issue in Section 5.

An important consideration is that we require a number of quantities in order to compute $\mathrm{E}_{\mathcal{G}}[y]$ via (4). In particular, we must either specify or compute $\mathrm{E}[y]$, $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$. Prior to offering practical suggestions for how this might be achieved, we note that our account thus far has presented an argument for why (4), if it can be computed, is closer to actual prevision and is thus important. If the meaning of any posterior statements that we will make about y having obtained data z is important, we have provided a route towards clear meaning through (4). Specifically, we have shown that $\mathrm{E}_{\mathcal{G}}[y]$ is as close as possible with respect to squared error loss to time t prevision of the true y. This represents a novel foundational and methodological step, focusing our attention on the further key quantities we require to claim ownership of posterior belief statements. As such, an interesting avenue for further research in this matter could focus on precisely how key quantities for posterior belief assessment should be obtained.

That said, we now discuss two particular ideas. The first is to focus elicitation on $\mathrm{E}[y]$, $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$. The elicitation of purely second order judgements is necessarily simpler than the sort of careful elicitation required for a fully subjective Bayesian analysis in a complex problem. This is discussed in many sources, an overview is in Chapter 2 of Goldstein and Wooff (2007). For example, though an expert may find it extremely hard to think about distributions for hyper-parameters of distributions on other parameters that purportedly control the distribution of some quantity related to the complex system y (wherein her expertise lies) through a conditional probability model, she will be far more comfortable considering $\mathrm{E}[y]$ and $\mathrm{Var}[y]$ as it involves thinking about the quantity she understands directly. Though simpler than a fully probabilistic elicitation exercise for the statistical model in question, the above argument may not be viewed, in certain problems, to be as strong when it comes to determining those key judgements involving $\mathcal{G}$. Hence we provide a second suggestion involving sampling.

4.2 An algorithm for computing posterior belief assessments

Our algorithm for obtaining $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$, the key quantities required to compute (4), requires that $\mathrm{E}[y]$ and $\mathrm{Var}[y]$ are specified, either through elicitation or otherwise. We can then form a distribution $(\mathrm{E}[y], \mathrm{Var}[y])$ and sample values of y. We also use the same arguments to form a distribution for the observables using the mean and variance $(\mathrm{E}[z|y], \mathrm{Var}[z|y])$. For example, z might be a measurement on y made with mean zero, uncorrelated error with given variance σ_e^2 and $\mathrm{E}[z|y] = y$ and $\mathrm{Var}[z|y] = \sigma_e^2$. We can now sample a value of y from $(\mathrm{E}[y], \mathrm{Var}[y])$ and corresponding values of z using $(\mathrm{E}[z|y], \mathrm{Var}[z|y])$. Once we are able to generate such samples, our algorithm proceeds as follows.

Step 1 Sample a value $\hat{y}$ and $\hat{z}$ from $(\mathrm{E}[y], \mathrm{Var}[y])$ and $(\mathrm{E}[z|\hat{y}], \mathrm{Var}[z|\hat{y}])$, respectively.

Step 2 Use the full Bayes machinery to compute each $\mathrm{E}[y|\hat{z}; J_0]$, $\mathrm{E}[y|\hat{z}; J_1], \ldots$, and use them to form $\hat{\mathcal{G}}$.

Step 3 Repeat this process to obtain a large number, N, of sample pairs $(\hat{y}_1, \hat{\mathcal{G}}_1), \ldots, (\hat{y}_N, \hat{\mathcal{G}}_N)$.

Step 4 Assess $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$ by computing the sample means and variances of the $\hat{\mathcal{G}}s$ and their covariance with the $\hat{y}$s.

To make the sampling algorithm computationally feasible using distributed computing, we may have to, for example, reduce the number of samples in our MCMC at step 2. When we do not have a firm judgement as to our choice for the distribution of y or $z|y$, then, for negligible computational cost, we may repeat the analysis under a variety of choices, by re-weighting the samples that we obtain using our algorithm, under each such choice.

Performing such alternative analyses by re-weighting may be important in the case where we have no firm choice for these distributions or where the model we use to obtain

$z|y$ is tentative or one of a collection of models we are comfortable with. Depending on the problem, there is a potential for the impact on the posterior belief assessment of these choices to be non-negligible. If, through re-weighting to other reasonable forms of distribution for y and $z|y$, we establish this to be the case, we would invest more time and effort eliciting features (such as skew, kurtosis, certain percentiles of the distribution function) of these distributions from our experts. As these quantities (each of which is an expectation) are functions of the observables, we view this as easier for a domain expert than elicitation of parameter distributions in our complex statistical models. However, we do not require that any elicited distribution here represent the subjective beliefs of the analyst or expert in the sense that every or any probability statement made is believed, as we are using probability here in the way described near the beginning of Section 4: as a tool to assist with obtaining a handful of beliefs we are prepared to adopt as prevision through the machinery of posterior belief assessment. The impact of these choices on the results of our sampling algorithm and the posterior belief assessment itself, as well as formal methods for assessing this and handling large impacts would be an interesting avenue for future research in this area.

4.3 Comparison with robust Bayes and sensitivity analysis

A robust Bayesian analysis requires us to be able to explore, analytically (usually), the behaviour of any key posterior quantities under all possible classes of prior and likelihood combination of interest. This is an extremely complex and arduous task, and may be infeasible for complex models. Even if completed, however, unless a particular posterior quantity is (practically) invariant to all choices, it is unclear what belief statements you can make and how they might relate to your prevision. Similarly, though a sensitivity analysis has the potential to be just as thorough as a posterior belief assessment in terms of the alternative models and priors explored ($J_1, \ldots, J_k$ in our account thus far), it leaves any formal relationship between the quantities computed under alternative judgements, your beliefs and prevision unestablished. Typically, we might report posterior beliefs under J_0 and use the alternative analyses to comment on sensitivity of these beliefs to our judgements, without actually incorporating data from any further analyses into our beliefs given the data.

A posterior belief assessment offers a practical and tractable investigation into the robustness of any conclusions you may wish to draw to alternative prior and likelihood choices. Though requiring extra computing power, the calculations required are essentially repeats of previous analyses and are therefore relatively straightforward to undertake, regardless of the complexity of the statistical modelling. Further, our approach reduces your uncertainty about the quantities of interest and gives an actual quantification of the current state of your beliefs about each quantity that accounts for all of the additional information contained in further analyses, through $\mathrm{E}_{\mathcal{G}}[y]$.

The account in this section has described posterior belief assessment for the situation where we can write down and construct full Bayes analyses for a finite set of alternative judgements $J_0, J_1, \ldots, J_k$. However, in many practical cases of interest, it is unlikely that k will be small enough to permit an exhaustive analysis, or perhaps the number of not ruled out judgements is, in principle, infinite (for example, we may have had to specify

continuous hyper parameters). In the next section, we outline how exchangeability may be imposed on our judgements and exploited to provide posterior belief assessment in more realistic and general situations.

5 Co-exchangeable classes and posterior belief assessment

One of the key tools in the subjective Bayesian's armoury when faced with a collection of similar random quantities is the judgement of exchangeability. Suppose we have collection of separate flips of a coin, $X_1, X_2, \ldots$, and we want to predict the outcome of some future flip of the same coin. The judgement of exchangeability, effectively that our judgements for any combination of the Xs does not depend on the indices, allows us to treat each X_i as a draw from some probability distribution, to learn about the parameters of that distribution using our collection of flips, and to make inference about future, yet unobserved Xs. This is a consequence of the representation theorem (de Finetti, 1974), and a detailed introductory discussion is given in Goldstein (2012).

Having a collection of alternative judgements $J_0, J_1, \ldots$ leading, in principle, to quantities $\mathrm{E}[y|z; J_0]$, $\mathrm{E}[y|z; J_1], \ldots$ leads us to ask questions such as are there any exchangeability judgements we might be willing to make and what might they be? There is no reason to assume, in general, that the collection $\mathrm{E}[y|z; J_0], \mathrm{E}[y|z; J_1], \ldots$ should be exchangeable. For example, the way J_i affects either the modelling or the prior choices might be quite different from another J_j, and this might lead us to strong a priori views on the differences between $\mathrm{E}[y|z; J_i]$ and $\mathrm{E}[y|z; J_j]$.

The key idea in this section will be to group our alternative judgements into classes within which we are prepared to impose some form of exchangeability on $\mathrm{E}[y|z; J_{i_1}]$, $\mathrm{E}[y|z; J_{i_2}], \ldots$ for $J_{i_1}, J_{i_2}, \ldots$ in class i. This will allow us to perform a handful of alternative Bayesian analyses for some of our not ruled out judgements and to use the information from these, to learn about other analyses, for example, those that we do not have time or the ability to perform. Before detailing our approach, we must first set up the type of exchangeability we require, namely *second order exchangeability*.

5.1 Second order exchangeability

In order to establish an inferential framework within which alternative Bayesian analyses may be used to discover relationships between explored judgements and the rest of our not ruled out judgements, we require further statistical modelling. We prefer to minimise the number of additional judgements required in order to do this, and so turn to second order exchangeability (SOE). A collection of quantities $X_1, X_2, X_3, \ldots$ are SOE if

$$\mathrm{E}[X_i] = \mu; \qquad \mathrm{Var}[X_i] = \Sigma \quad \forall i; \qquad \mathrm{Cov}[X_i, X_j] = \Gamma \quad \forall i \neq j.$$

In words, SOE represents an invariance of our judgements about means, variances and pairwise covariances in a collection to labelling. For an infinite SOE collection, the

representation theorem (Goldstein and Wooff, 2007, chapter 6) gives an orthogonal decomposition of each X_j into a common term $\mathcal{M}(X)$ and a residual $\mathcal{R}_j(X)$:

$$X_j = \mathcal{M}(X) + \mathcal{R}_j(X) \tag{5}$$

with

$$\mathrm{E}\left[\mathcal{M}(X)\right] = \mu; \qquad \mathrm{Var}\left[\mathcal{M}(X)\right] = \Lambda; \tag{6}$$

$$\mathrm{E}\left[\mathcal{R}_j(X)\right] = 0; \qquad \mathrm{Var}\left[\mathcal{R}_j(X)\right] = \Sigma - \Lambda; \tag{7}$$

and

$$\mathrm{Cov}\left[\mathcal{M}(X), \mathcal{R}_j(X)\right] = \mathrm{Cov}\left[\mathcal{R}_j(X), \mathcal{R}_k(X)\right] = 0 \qquad \forall\, j, \text{and } k \neq j. \tag{8}$$

The inferential power of this judgement is that it leads to the orthogonality of a set of n observations $D_n = (X_1, X_2, \ldots, X_n)$ and any X_j, $j > n$, given $\mathcal{M}(X)$. This orthogonality, termed *Bayes linear sufficiency*, means that the Bayes linear update $\mathrm{E}_{D_n \cup \mathcal{M}(X)}\left[X_j\right]$, is equivalent to $\mathrm{E}_{\mathcal{M}(X)}\left[X_j\right]$, which implies (cf. Goldstein and Wooff, 2007, page 195)

$$\mathrm{E}_{D_n}\left[X_j\right] = \mathrm{E}_{D_n}\left[\mathcal{M}(X)\right]. \tag{9}$$

Hence, suppose we judged that $\mathrm{E}\left[y|z; J_1\right]$, $\mathrm{E}\left[y|z; J_2\right], \ldots$, were an infinite SOE sequence and that we only observed $D_k = \{\mathrm{E}\left[y|z; J_k\right]; k = 1, \ldots, 4\}$, then any $\mathrm{E}_{D_k}\left[\mathrm{E}\left[y|z; J_i\right]\right]$, $i > 4$, is equal to $\mathrm{E}_{D_k}\left[\mathcal{M}\right]$ and can be computed using (5), (6), (7) and (8). (Note that a Bayes linear update takes the same form as (4).)

In general, we will not wish to impose SOE across the collection of posterior expectations under all alternative judgements we might make, as we may have relatively strong judgements regarding the relationship between certain alternative modelling choices. For example, if J_i and J_j differ only in the values of the hyper-parameters of one particular prior distribution, and J_k represents an alternative formulation of the likelihood, then we might view, a priori, that

$$\mathrm{Cov}\left[\mathrm{E}\left[y|z; J_i\right], \mathrm{E}\left[y|z; J_j\right]\right] \neq \mathrm{Cov}\left[\mathrm{E}\left[y|z; J_i\right], \mathrm{E}\left[y|z; J_k\right]\right]$$

and similarly for $\mathrm{Cov}\left[\mathrm{E}\left[y|z; J_j\right], \mathrm{E}\left[y|z; J_k\right]\right]$. The potential existence of such relationships between posterior conditional expectations under different judgements motivates a partitioning of our judgements into co-exchangeable classes.

5.2 Co-exchangeable classes

Let $Y_1, Y_2, \ldots$ be a collection of sequences. We say that they are *co-exchangeable* if, for any fixed k, the sequence $Y_{k1}, Y_{k2}, \ldots$ is infinite second order exchangeable and if, for $j \neq k$, $\mathrm{Cov}\left[Y_{jm}, Y_{kn}\right] = \Sigma_{jk}$ for any m, n. Let our alternative judgements be partitioned into k classes $J_1, \ldots, J_k$ with $J_i = \{J_{i_1}, J_{i_2}, \ldots\}$, each chosen so that we may define corresponding co-exchangeable classes $C_1, \ldots, C_k$ with $C_i = \{\mathrm{E}\left[y|z; J_{i_1}\right], \mathrm{E}\left[y|z; J_{i_2}\right], \ldots\}$. Then, the representation theorem gives

$$\mathrm{E}\left[y|z; J_{i_j}\right] = \mathcal{M}(C_i) + \mathcal{R}_j(C_i). \tag{10}$$

Equation (10) and co-exchangeability between classes means that

$$\mathrm{Cov}\left[\mathrm{E}\left[y|z;J_{i_j}\right],\mathrm{E}\left[y|z;J_{m_l}\right]\right]=\mathrm{Cov}\left[\mathcal{M}(C_i),\mathcal{M}(C_m)\right]\qquad i\neq m, \tag{11}$$

and, from (6), (7) and (8), we have

$$\mathrm{Cov}\left[\mathrm{E}\left[y|z;J_{i_j}\right],\mathrm{E}\left[y|z;J_{i_l}\right]\right]=\mathrm{Var}\left[\mathcal{M}(C_i)\right]\qquad j\neq l. \tag{12}$$

By partitioning into co-exchangeable classes in this way, we are able to run a small number of alternative Bayesian analyses corresponding to a subset of alternative judgements, to use these to update beliefs about any other Bayesian analysis we might have performed and to compute our posterior belief assessment. Suppose we were to compute posterior expectations based on Bayesian analyses on N sets of alternative judgments. Let $n(1),\ldots,n(N)\in\{1,\ldots,k\}$ and $m(1),\ldots,m(n)\in\mathbb{Z}$ be indices corresponding to the chosen judgements, so that the jth Bayesian analysis performed happens under $J_{n(j)_{m(j)}}$. Define $D=\{\mathrm{E}[y|z;J_{n(1)_{m(1)}}],\ldots,\mathrm{E}[y|z;J_{n(N)_{m(N)}}]\}$, then (11) and (12) lead to a version of (9), namely

$$\mathrm{E}_D\left[\mathrm{E}\left[y|z;J_{i_j}\right]\right]=\mathrm{E}_D\left[\mathcal{M}(C_i)\right]. \tag{13}$$

Further, each $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$, for $i=1,\ldots,k$, is linear combination of the elements of D. Hence, when looking to use the elements of D to perform a posterior belief assessment, we can simultaneously include D and all of the information contained therein regarding unobserved alternative Bayesian analyses by selecting $\mathcal{G}$ to be the vector $(\mathrm{E}\left[y|z;J_0\right],\mathrm{E}_D\left[\mathcal{M}(C_1)\right],\ldots,\mathrm{E}_D\left[\mathcal{M}(C_k)\right])=(\mathcal{G}_1,\ldots,\mathcal{G}_k)$. Letting $\mathcal{G}_0$ be the unit constant, then the results in Theorem 1 hold.

5.3 Special case: co-exchangeability of y

A further co-exchangeability judgement simplifies the task of prior specification required in order to compute $\mathrm{E}_{\mathcal{G}}\left[y\right]$: namely, that of co-exchangeability between y and the members of each class, $C_1,\ldots,C_k$.

Determining $\mathrm{Cov}\left[y,\mathcal{G}\right]$ requires determining $\mathrm{Cov}\left[y,\mathrm{E}\left[y|z;J_0\right]\right]$ and $\mathrm{Cov}\left[y,\mathrm{E}_D\left[\mathcal{M}(C_i)\right]\right]$ for $i=1,\ldots,k$. Supposing we have the quantities required to compute each $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$, then the co-exchangeability of y and the members of each class $C_1,\ldots,C_k$ means that we require only $\mathrm{Cov}\left[y,\mathrm{E}\left[y|z;J_0\right]\right]$ and $\mathrm{Cov}\left[y,\mathcal{M}(C_i)\right]$ for $i=1,\ldots,k$ in order to compute $\mathrm{Cov}\left[y,\mathcal{G}\right]$.

To show this, we first note that

$$\begin{aligned}\mathrm{Cov}\left[y,\mathrm{E}_D\left[\mathcal{M}(C_i)\right]\right]&=\mathrm{Cov}\left[y,\mathrm{Cov}\left[\mathcal{M}(C_i),D\right]\mathrm{Var}\left[D\right]^{-1}D\right]\\&=\mathrm{Cov}\left[y,D\right]W_i\end{aligned}$$

with $W_i=\mathrm{Var}\left[D\right]^{-1}\mathrm{Cov}\left[D,\mathcal{M}(C_i)\right]$. W_i is required in order to compute $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$, hence, the additional burden in prior specification in order to compute $\mathrm{Cov}\left[y,\mathrm{E}_D\left[\mathcal{M}(C_i)\right]\right]$ requires us to compute $\mathrm{Cov}\left[y,D\right]$. From (10),

$$D_j=\mathcal{M}(C_{n(j)})+\mathcal{R}_{m(j)}(C_{n(j)}).$$

Defining $\tilde{\mathcal{M}} = \mathcal{M}(C_{n(1)}), \ldots, \mathcal{M}(C_{n(N)})$, then co-exchangeability of y implies that

$$\mathrm{Cov}\left[y, D\right] W_i = \mathrm{Cov}\left[y, \tilde{\mathcal{M}}\right] W_i$$

so that we need only specify $\mathrm{Cov}\left[y, \mathcal{M}(C_i)\right]$ for $i = 1, \ldots, k$ in order to compute $\mathrm{Cov}\left[y, \mathrm{E}_D\left[\mathcal{M}(C_i)\right]\right]$. If we are unwilling or unable to specify prior covariances here, we may obtain the $\mathrm{Cov}\left[y, \mathrm{E}_D\left[\mathcal{M}(C_i)\right]\right]$ using a similar sampling scheme to that described in Section 4.2.

6 Application: calibrating an expensive ocean model

We return to the ocean model calibration problem described in Section 2, with the model established by Figure 1. We first describe the ocean model and the data we will use to calibrate it before returning to the problem of establishing judgements for prior and likelihood that will enable the formal Bayesian analysis to proceed. We will also describe a collection of alternatives and perform the posterior belief assessment using the methodology described in Sections 4 and 5.

6.1 The ocean model and the data

The NEMO ocean model (Madec, 2008) simulates the global ocean for given atmospheric forcing files. The 2° ORCA configuration that we use here divides the globe into approximately 2° latitude/longitude grid boxes and has 32 depth levels. The model is extremely computationally expensive, taking approximately 7.5 hours to complete 30 years of model time on the UK supercomputer ARCHER. For any setting of the parameters, we run the model for 180 model years under climatological forcing (atmospheric forcing designed to replicate an "average year" in the late 20th century), and perform our analysis on the results of the last 30 years of the simulation.

The initial parameter space, containing 1 switch variable with 2 settings and 20 continuous parameters defined on a hypercube with the ranges for each input ($a_{x^*}^{[i]}$ and $b_{x^*}^{[i]}$ in our Bayesian network) elicited by the first author from the developer of the code, Gurvan Madec, is explored using a 400 member k-extended Latin Hypercube design (LHC), a Latin Hypercube of size 400 constructed based on an initial LHC and extended $k-1$ times so that the design comprises k smaller LHCs of size $400/k$ (in this case k was 25). The design method and rationale for this choice applied to the 2° NEMO model is discussed in Williamson (2014). Figure 2 plots global mean temperature as a function of depth with cyan lines representing the 400 model runs and the red solid line representing real world data (using the EN3 dataset, Ingleby and Huddleston, 2007). Dashed red lines represent ±2 standard deviations of the model error calculated from Ingleby and Huddleston (2007). The dashed horizontal lines represent the 31 model depths.

To illustrate a posterior belief assessment, we will calibrate the 2° NEMO using observations at 4 depths (illustrated by the black dashed lines) in order to predict

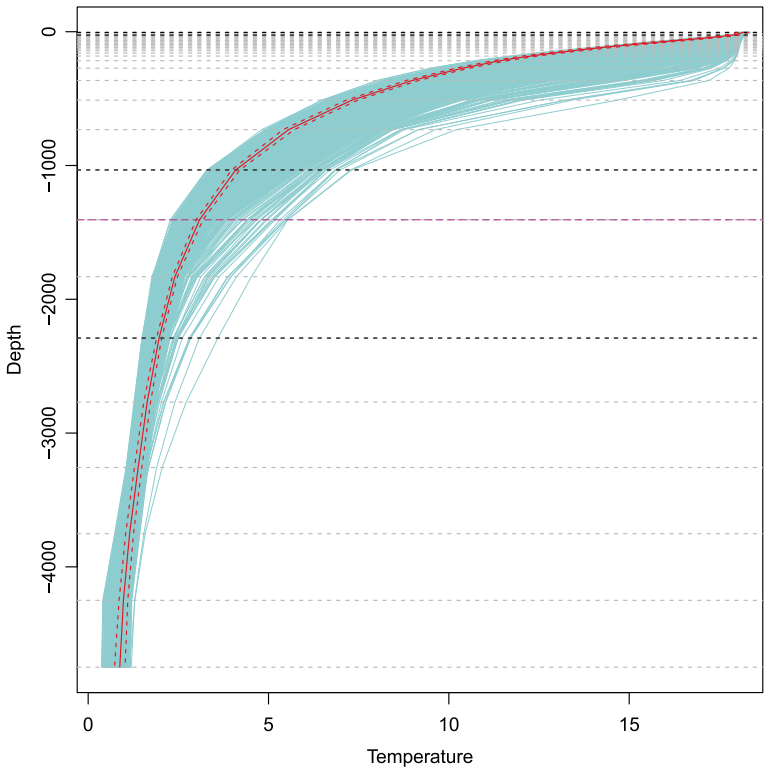

Figure 2: Global mean temperatures for each of the 400 members of our ocean model ensemble (cyan lines) plotted against model depth. The red solid line represents observed global mean temperature and the dashed red lines indicate ±2 standard deviations of the error on the observations. Horizontal dashed lines indicate the model depth levels, with the black dashed lines those depths used to calibrate the model and the magenta long-dashed line the prediction depth.

the temperature at a new depth for which we pretend not to have observed data (the magenta long-dashed line). Though clearly illustrative (the goal of this paper is not to provide a detailed ocean reanalysis based on NEMO), this analysis will be sufficiently complex to highlight the features of our methodology and to provide a roadmap for posterior belief assessment in the calibration of computer models.

6.2 Alternative judgements

Even with the detailed modelling in Section 2, there is a great deal left to specify before a Bayesian analysis can be completed. The remaining choices for prior and likelihood are part of one member of a collection of judgements from $J_0, J_1, \ldots$. We describe the elements of J_0 and list the types of possible alternatives to each element we consider

and the rationale behind choosing them. We will then divide the types of alternatives into a number of co-exchangeable classes, following Section 5, and select a number of representatives for each class.

We begin with judgements concerning the prior and likelihood of the emulator component, starting with the type and number of terms in response surface component $g_d(x)$. Our preference is to fit complex response surfaces with $g_d(x)$ containing complex polynomial terms in x chosen using some model selection algorithm. There is considerable literature adopting this approach (cf. Craig et al., 2001; Cumming and Goldstein, 2009; Williamson et al., 2012; Vernon et al., 2010; Kaufman et al., 2011; Williamson et al., 2014), which is based on a philosophy that global effects can be captured using the response surface, allowing local effects to be captured by the Gaussian process residual. We use a stepwise selection routine described in Williamson et al. (2013) similar to that used in Sexton et al. (2011) that aims to add as many variance describing terms to each $g_d(x)$ as possible and then deletes those that describe the least variance, one at a time, until there are fewer terms than a maximum allowed number (specified as a percentage of the number of degrees of freedom available) and the explained variance of the best current best candidate for deletion is above a certain threshold (usually 0.1%). Our default choice is to use 10% of the available degrees of freedom allowing up to 40 terms in each $g_d(x)$. This is part of J_0.

As well as the potential to choose fewer terms using our model selection routine, we also consider other approaches that are popular within the literature. A popular choice is $g_d(x) = (1, x_1, \ldots, x_r)^T$ (Andrianakis and Challenor, 2012; Lee et al., 2013), as is the choice made in the original computer experiments paper of $g_d(x) = 1$ (Sacks et al., 1989). We view these two as not ruled out alternatives reflecting, for example, a difference in expert opinion in the form of the likelihood. In this study, as in most applications with emulators (cf. Haylock and O'Hagan, 1996; Kennedy and O'Hagan, 2001; Santner et al., 2003), we use the reference prior for response surface coefficients so that $\pi(\beta_d, \sigma_f^2) \propto 1/\sigma_f^2$.

Popular choices of correlation function $R(|x-x'|; \kappa_{1:r})$ in the literature include power exponential forms and Matérn forms. Power exponential correlation functions take the form

$$R(|x - x'|; \kappa_{1:r}) = \prod_{i=1}^{r} \exp\{-\kappa_i |x_i - x_i'|^p\}$$

with the special case $p = 2$ being the Gaussian covariance function. Though this is popular, Bayarri et al. (2007) suggested $p = 1.9$ was a better alternative as the residual under the Gaussian correlation function was often too smooth; and others favour Matérn forms, which are popular in spatial statistics. The two Matérn functions we consider are chosen for computational convenience and have

$$R(|x - x'|; \kappa_{1:r}) = \prod_{i=1}^{r} (1 + \sqrt{3}\kappa_i |x_i - x_i'|) \exp\{-\sqrt{3}\kappa_i |x_i - x_i'|\}$$

and

$$R(|x - x'|; \kappa_{1:r}) = \prod_{i=1}^{r} (1 + \sqrt{5}\kappa_i |x_i - x_i'| + \frac{1}{3}\sqrt{5}\kappa_i |x_i - x_i'|^2) \exp\{-\sqrt{5}\kappa_i |x_i - x_i'|\}.$$

In J_0, we select the power exponential form with $p = 1.9$ as this is the form we are most familiar with. The other forms (including powers p closer to 1) change both the prior and the likelihood and are considered as alternatives.

We used the MATCH elicitation tool (Morris et al., 2014) to select values for the hyper-priors $a_\kappa^{1:r}$ and $b_\kappa^{1:r}$, selecting values of 2.9 and 5, respectively, for the hyper parameters in each dimension for J_0. We do not rule out alternatives with the same prior mean for each κ, but allow the variance to be up to twice as large or as low as 5 times smaller. As the nugget in this paper represents the proportion of the residual variability that is due to the initial condition uncertainty in the model, we base our prior judgements for ν on how much of the signal in the data we believe our response surface will capture. Given $g_d(x)$ the emulator is fitted using the joint Bayesian update described in Haylock and O'Hagan (1996). However, we select $g_d(x)$ using a stepwise selection routine (described above) based on ordinary least squares. Following selection of $g_d(x)$, we use the R^2 value of the last fit in the stepwise routine to judge how much of the residual variability should be uncorrelated based on 7 separately elicited scenarios. For example, the first scenario is that the OLS fit captures more than 95% of the variability in the ensemble. In this case we believe the residual is mainly uncorrelated. We use the MATCH tool to elicit a $\text{Be}(3.8, 1.7)$ distribution for ν that is skewed towards high values. The other scenarios are $R^2 > 0.9$, $R^2 > 0.8$, $R^2 > 0.7$, $R^2 > 0.6$, $R^2 > 0.4$ and $R^2 \leq 0.4$, with MATCH elicited priors $\text{Be}(2.3, 1.7)$, $\text{Be}(2, 1.5)$, $\text{Be}(1.5, 1.5)$, $\text{Be}(1.6, 2.1)$, $\text{Be}(1.8, 3.8)$ and $\text{Be}(1.4, 3.1)$, respectively. These judgements are in J_0. We consider alternative judgements that respect the mean of distribution for ν but inflate the variance by multiplying both a_ν and b_ν by a scalar as large as 4.

As we have σ_e^2 (as shown by the red dashed lines on Figure 2), all that remains is to specify hyper-parameters controlling the model discrepancy. For ζ we choose a prior for the roughness length that reflects our belief that if the model is biased at one depth, it is very likely to be similarly biased at nearby depths. We specify $a_\zeta = 1$, $b_\zeta = 7$ to reflect this. We do not explore changes to these judgements, however, we concentrate on the impact of specifying different discrepancy variance distributions. Our current judgements for this are $a_\eta = 1000$, $b_\eta = 6.8 \in J_0$. This gives an expectation for σ_η^2 equivalent to the observation error and is a value we have considered as the modeller's tolerance to error in our work with NEMO (see Williamson et al., 2015, for discussion).

Retaining the same mean value, we consider two alternative types of discrepancy judgement. The first type we describe as "medium" discrepancy where the hyper-parameters are in the same proportion but roughly an order of magnitude smaller so that the distribution on σ_η^2 has a slightly larger variance and is a little more skewed to higher values. The second is "large" discrepancy, which is similar, but with 2 orders of magnitude difference in the values of a_η and b_η so that σ_η^2 has a much larger variance and is very highly skewed towards higher values. Unlike some of the alternatives to the likelihood described earlier in this subsection, these alternatives represent quite different perspectives on the relevance of the NEMO ocean model to the actual ocean (for global mean temperature at least). Hence, we might view these alternatives as representative of different levels of confidence held by different ocean modellers about the NEMO model.

Before describing our co-exchangeable classes of alternative judgements, it is worth making a brief comment about the way the emulators are fitted and the calibration

performed. Though we do run MCMC to obtain posterior judgements about $y(d)$, we follow the standard approach in calibration and first fix the correlation and nugget parameters after conditioning on the ensemble (Kennedy and O'Hagan, 2001). This means first fitting the emulator, then using it in the calibration to the data. We choose the MAP estimates of $\kappa_{1:r}$ and ν instead of maximum likelihood estimates, as these account for our prior modelling. MAP estimates are obtained using simulated annealing.

6.3 Co-exchangeable classes of not ruled out alternatives

We will be interested in $\mathrm{E}[y(d_5)|z(d_{1:4})]$, the expected global mean temperature at around 1405 metres depth given the four observations we have taken. For many of the alternative modelling and prior choices we have described there may well be an impact upon this posterior expectation, however, due to the complexity of the model it is not clear to us, a priori, how any impacts might be related. However, we do have views regarding how the type of response surface we fit and how the different discrepancy choices might lead us to quite different analyses. Hence our division of the collection of possible alternative posterior expectations under different judgements will be based on these 2 parts of our modelling.

Considering the type of response surface fitted first, we view the alternatives as forming 3 distinct impacts upon our analysis. In the case of complex mean functions fitted using our stepwise routine, we consider that much of global signal will be captured by $\beta^T g_d(x)$, allowing the Gaussian process $\epsilon_d(x)$ to capture only local departures from this surface. Hence posterior estimates of σ_f^2, ν and $\kappa_{1:r}$ are likely to be much different than in the cases where either a linear or constant response surface is fitted. In the case of complex mean functions, as the $\epsilon_d(x)$ process will be more tuned to local deviations from the response surface rather than the behaviour of the function in the whole model parameter space, we believe it will be more accurate than the other emulators in regions of parameter space that are close to our design points. If we have any design points that are quite close to x^*, we expect $\mathrm{E}[y(d_5)|z(d_{1:4})]$ to be captured more accurately using this emulator than the other types.

However, we have no strong views on how reducing the number of allowed degrees of freedom to be spent on fitting the stepwise regression will impact upon this argument. As we only allow a minimum of 5% of the degrees of freedom to be spent on the response surface (in this case the same number of degrees of freedom spent using the linear response surface with all of the parameters), we believe that the major global non-linearities will be captured by $\beta^T g_d(x)$ anyway. Therefore, even though $\epsilon_d(x)$ may contain more of the global signal than if a higher number of degrees of freedom were spent, it may not (some of the additional terms retained in $\beta^T g_d(x)$ when a higher percentage of our degrees of freedom are spent may in fact be capturing local signal through high order interactions between inputs).

We view the cases of the linear (in the model inputs) and constant response surfaces to have distinct impacts upon our posterior expectation, both from the complex mean case (for the reason given above) and from each other, as we now explain. In the case of the emulator with linear mean, the posterior uncertainty of the emulator will increase

asymptotically as we move outside of the convex hull of design points (as it would with the complex mean), however, the same is not true of the constant mean emulator. Hence, we may view this increased uncertainty outside of the convex hull of design points as contributing to our posterior expectation in some way. Perhaps, for example, x^* is outside of that convex hull, leading to our range of possible values of the computer model at x^* to be quite different in the two cases. Or perhaps the model outputs that are possible in the linear mean case, but not in the constant mean case, bias the posterior expectation in one particular direction (perhaps they allow much higher temperatures, but not lower).

So far, we have defined 3 distinct classes of judgements based only on our choice of emulator. The choice of discrepancy variance parametrisation ("standard", "medium" or "high") will also impact upon our prior judgements for $\mathrm{E}[y|z]$. High values of model discrepancy variance reduce the impact of information from the ocean model on our posterior expectations for ocean temperature. If the discrepancy is high enough, it can effectively "take over" as there is little information in the data about the ocean model parameters so that our prediction is driven by the Gaussian process discrepancy fitted as (approximately) a deviation from the ensemble mean. We do not view the "medium" case as being different enough from the "standard" to change our beliefs about $\mathrm{E}[y|z]$. The medium case has a wider range for σ_η^2 and is slightly more skewed towards higher values, but only slightly and not enough to give us any strong views. However, we view the "high" case, which is much more skewed and allows much higher discrepancy variances, to be distinct from these other two.

Combinations of our two distinct classes driven by discrepancy judgements and 3 distinct classes based on the type of emulator lead to 6 classes of potential judgements. Within each class we have all choices of covariance function and hyper-priors for the emulator residual. By just considering our alternative hyper prior choices $a_\kappa^{1:r}$ and $b_\kappa^{1:r}$, (any values such that the prior mean of the half length correlation is the same but that the variance can be up to twice as large or as low as 5 times smaller), we can see that there are infinitely many alternative judgements within each class ($a_\kappa^{1:r}$ and $b_\kappa^{1:r}$ are continuous). Within each class we have no a priori views on how the means, variances or covariances of $\mathrm{E}[y|z]$ would differ with each possible alternative. We also have no views on any specific member dependent differences between $\mathrm{E}[y|z]$ calculated from different classes. Hence we assume that the 6 defined classes $C_1, \ldots, C_6$ are co-exchangeable.

6.4 Posterior belief assessment

Let J_{i_j} represent a collection of modelling judgements from C_i, then (10) gives

$$\mathrm{E}\left[y(d_5)|z(d_{1:4}); J_{i_j}\right] = \mathcal{M}(C_i) + \mathcal{R}_j(C_i),$$

and we can apply (6), (7), (8), (11) and (12) to compute $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$ for $i = 1, \ldots, 6$, with $D = \{\mathrm{E}[y(d_5)|z(d_{1:4}); J_{n(1)_{m(1)}}], \ldots, \mathrm{E}[y(d_5)|z(d_{1:4}); J_{n(N)_{m(N)}}]\}$ and with $J_{n(1)_{m(1)}}, \ldots, J_{n(N)_{m(N)}}$ chosen modelling choices from each class. Defining $\mathcal{G}$ as in Section 5, then in order to perform the posterior belief assessment and compute $\mathrm{E}_{\mathcal{G}}\left[y(d_5)\right]$ via (4) we require $\mathrm{E}\left[y(d_5)\right]$, $\mathrm{Cov}\left[y(d_5), \mathcal{G}\right]$, $\mathrm{Var}\left[\mathcal{G}\right]$, $\mathrm{E}\left[\mathcal{G}\right]$ and a means to computing $\mathcal{G}$

from D which, further, requires $\mathrm{E}[\mathcal{M}(C_i)]$, $\mathrm{Cov}[\mathcal{M}(C_i), \mathcal{M}(C_j)]$ and $\mathrm{Var}[\mathcal{R}_j(C_i)]$ for each $i = 1, \ldots, 6$.

We first select the further Bayesian analyses we will do to obtain D, by choosing a representative set of judgements from each class. We do this by first selecting a number of distinct alternative choices from those not ruled out in Section 6.3 for each prior/likelihood decision in our modelling. These are as follows: for classes involving complex response surfaces, we allow both 10% and 5% of the degrees of freedom to be spent in model selection. For the "medium" and "high" discrepancy settings we divide both the standard choices of the hyper-parameters by 10 and 100, respectively. For all classes, we allow both classes of Matérn, the Gaussian covariance function and the power exponential with $p = 1.5$. We multiply all hyper-parameters of $\kappa_{1:r}$ by 5 or 0.5, respectively, and allow the nugget hyper-parameters to be multiplied by 4. Each combination of alternative choices and our standard choices is also allowed.

Instead of running alternative Bayesian analyses for each possible combination, we take advantage of co-exchangeability, sampling 32 choices from C_1 (the largest class) and 8 from each of the remaining classes. Each Bayesian analysis builds the emulator determined by our modelling choices, then uses a random walk metropolis hastings algorithm with $21,000$ samples, discarding 1000 as burn in and thinning every 20, to obtain a sample from the posterior distribution implied by our modelling choices from which an expectation can be computed. This gives D, however, in order to compute $\mathcal{G}$ and $\mathrm{E}_{\mathcal{G}}[y(d_5)]$ we require each of the other ingredients described above. We obtain these here by sampling.

We use the sampling scheme described in Section 4.2. The first step of this process was to be able to form distributions $(\mathrm{E}[y], \mathrm{Var}[y])$ and $(\mathrm{E}[z|y], \mathrm{Var}[z|y])$ and to be able to sample values of y then z. Given y, we can easily obtain z using σ_e^2 and our statistical model. In an application where our judgements would be crucial in informing scientists about global temperatures in the real ocean, rather than one conducted for illustrative purposes, we would prefer to attempt to elicit means and variances for y directly. However, as this example is illustrative, we do not have the expertise available to do this. We therefore sample values of y by sampling one of our sets of judgements at random and forming the emulator $\hat{f}(x)$, sampling an x^*, drawing a sample from $\hat{f}(x)$ at this sample and adding it to a sample from the discrepancy distribution.

Having generated a y and z pair, we run a similar MCMC calculation to that described above in order to generate a sample of D under the sampled y and z in the manner described in Section 4.2 step 2. To make things computationally more feasible we run shorter simulations (6000 MCMC steps) with the same number discarded as burn in, thinning every 10. We took 13,000 samples of the 72 elements of D for different y and z pairs like this using the condor cluster at Durham University Maths department to complete all calculations within a day.

We first use these samples to assess $\mathrm{E}[\mathcal{M}(C_i)]$, $\mathrm{Cov}[\mathcal{M}(C_i), \mathcal{M}(C_j)]$ and $\mathrm{Var}[\mathcal{R}_j(C_i)]$ for each $i = 1, \ldots, 6$. We obtain the first two of these three quantities for each class by taking the mean over all experiments for each member of D and by computing the mean and variances of these within each class. The latter is assessed by taking the mean

of the variances over all experiments for each class and subtracting the assessment for the variance of $\mathcal{M}(C_i)$. For simplicity, we assume $\mathrm{Cov}\left[\mathcal{M}(C_i), \mathcal{M}(C_j)\right] = 0$ for $j \neq i$ as each had a small variance so was treated as known. When conducting these samples, an important part of any application of our methodology to a real world problem would be to perform diagnostics, using the samples, on our exchangeability judgements. Whilst we do not have the space to address this here, we comment that the sampling method offers the chance to examine the validity of our statistical modelling without confronting it with our only real data.

For each sample, we now compute the sample Bayes linear adjusted expectation $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$ using the quantities derived as above, and use them to form sample values of $\mathcal{G}$ as in step 3 of our algorithm. We now have coherent sample values of y and $\mathcal{G}$ and use them to establish all quantities required to compute $\mathrm{E}_{\mathcal{G}}\left[y\right]$ via (4) as described in step 4. We then compute the actual $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$ based on our observation driven MCMC calculations, form $\mathcal{G}$ and compute $\mathrm{E}_{\mathcal{G}}\left[y\right]$. The values of $\mathrm{E}_{\mathcal{G}}\left[y(d_5)\right]$ and $\mathrm{E}\left[y(d_5)|z(d_{1:4}); J_0\right]$ are 2.921°C and 2.951°C, respectively. The adjusted variance in y following each analysis, $\mathrm{Var}\left[y(d_5)\right] - \mathrm{Var}\left[\mathrm{E}_{\mathcal{G}}\left[y(d_5)\right]\right]$ and $\mathrm{Var}\left[y(d_5)\right] - \mathrm{Var}\left[\mathrm{E}_{\mathcal{G}_1}\left[y(d_5)\right]\right]$ was 0.0262 and 0.0226. As described in Section 4, part (i) of the theorem implies that a lower bound on the amount of uncertainty reduction in $y(d_5)$ achieved performing a posterior belief assessment (computed by taking the ratio of these two adjusted variances) is 14.0%.

We note here that our choice of how many samples to take from each class was largely arbitrary in this problem. However, if we built in a way of using elements of D to learn about $\mathrm{Var}\left[\mathcal{R}_j(C_i)\right]$ before then computing $\mathrm{E}_D\left[\mathcal{M}(C_i)\right]$ as part of what Goldstein and Wooff (2007) called a two stage Bayes linear analysis, we would be able to use this information to see how our uncertainty in $\mathcal{M}(C_i)$ was reduced using D. The idea would be to use information from this procedure to decide how many samples would be required from each class. Developing a two stage procedure for posterior belief assessment would be technically challenging and is a possible avenue of future research.

The posterior belief assessment for y is a constant plus a linear combination of the elements of $\mathcal{G}$. Hence it is instructive to observe what the coefficients of this linear combination are. We show these in Table 1. The larger coefficients for the three classes corresponding to the high discrepancy case indicate that this judgement may represent the main difference between our expectation following posterior belief assessment and that conditioned only on our original judgements. Though the goal of this analysis is not model selection, these coefficients do offer some a posteriori information about the relative merit of each class of model. If model selection was of interest, any particularly influential classes might be further explored by looking at the particular linear combination of the elements of D that influence its adjusted expectation to see if there were any particularly influential conditional expectations. We note, however, that the elements of $\mathcal{G}$ (and of D) are not orthogonal so any interpretation as to the relative merits of any class of model (or individual model) on the posterior belief assessment is not straightforward.

The adjusted variance, $\mathrm{Var}\left[\mathrm{E}_{\mathcal{G}}\left[y\right]\right]$ is such that $\mathrm{E}\left[y|z; J_0\right]$ is not even 1 standard deviation away from $\mathrm{E}_{\mathcal{G}}\left[y\right]$. Hence, in this example, our modified posterior judgements

are close to those from the initial Bayesian analysis under J_0. This is not strictly a robust Bayesian analysis, however, given the number of alternative modelling choices explored, we would view our conclusions as robust to both the prior and likelihood choices we have made. Though there are obvious similarities with sensitivity analysis here, we note that our inference actually combines all of the information from the sensitivity study, rather than only using the extra information as a sanity check. Further, if our analysis suggested, for example, that our assessment was sensitive to the samples we had taken, that would only suggest performing further analyses in order to reduce $\mathrm{Var}\left[\mathrm{E}_{\mathcal{G}}\left[y\right]\right]$, rather than requiring us to think more carefully about certain aspects of the prior or likelihood.

$\mathrm{E}\left[y\middle\vert z;J_0\right]$	$\mathrm{E}_D\left[\mathcal{M}(C_1)\right]$	$\mathrm{E}_D\left[\mathcal{M}(C_2)\right]$	$\mathrm{E}_D\left[\mathcal{M}(C_3)\right]$	$\mathrm{E}_D\left[\mathcal{M}(C_4)\right]$	$\mathrm{E}_D\left[\mathcal{M}(C_5)\right]$	$\mathrm{E}_D\left[\mathcal{M}(C_6)\right]$
0.133	0.383	−0.078	0.361	1.140	1.024	−0.842

Table 1: Coefficients from $\mathrm{E}_{\mathcal{G}}\left[y\right]$ for each of the elements of $\mathcal{G}$.

7 Discussion

Performing a truly subjective Bayesian analysis (where all elements of the prior modelling are held subjective beliefs, leading to a clear interpretation of the posterior distribution) in problems requiring complex statistical models, would be extremely challenging, expensive and time consuming. This would be the case even if elicitation of high-dimensional joint distributions, even for non-standard forms, was well supported and routine in Bayesian analysis. However, the case where we require posterior beliefs for every, or even very many, combinations of quantities in our statistical model is perhaps rare at best, though likely non-existent. Posterior belief assessment is a methodology that allows us to use the full Bayesian machinery in order to obtain beliefs for key quantities of interest using as many alternative forms of prior and likelihood modelling judgements as we are prepared to consider as representative of our judgements regarding the problem structure and our beliefs before undertaking the analysis.

We showed that, as a consequence of temporal sure preference, the posterior belief assessment is closer to the prevision we would specify at time t, than a full Bayesian analysis on our initial/default set of modelling judgements. We showed that we expect a posterior belief assessment to resolve more of our uncertainty about those key judgements of interest than a standard Bayesian analysis.

We argued that posterior belief assessment was a powerful and tractable alternative to traditional robust Bayesian analyses (that attempt to proceed analytically) when of interest is how robust any key judgements might be to plausible alternatives to modelling. Whilst a traditional analytic robust Bayesian analysis might not be possible in most applications with complex statistical models, we can perform a posterior belief assessment as long as we are able to repeat the Bayesian sampling computation (either in parallel or otherwise) for a finite set of alternative judgements. Our approach both gives us posterior judgements about quantities of interest, the difference between these and a one-off Bayesian analysis, and information as to which alternative prior modelling choices contribute to these and by how much.

Though tractable, particularly when compared to robust Bayes in complex statistical models, a posterior belief assessment requires a number of additional full Bayes analyses, that will likely involve computationally expensive sampling procedures to be repeated. If it would exhaust the analyst's computational resources to perform one of these analyses, posterior belief assessment will not be feasible. However, in most cases, either additional time or access to distributed computing facilities will be available and will allow the extra calculations to be done. In some cases, an alternative set of judgements may lead to a statistical model for which the Bayesian machinery has not yet been adequately developed to sample from, or that would involve a great deal of extra effort to implement. In these cases, it may be that such models belong to a co-exchangeable class containing models that are easier to sample from. If so, we run the simpler alternatives from this class and proceed as discussed in Section 5. If difficult to implement alternatives are not second order exchangeable with simpler alternatives, then further methodology may be required to perform a posterior belief assessment.

The principal challenge when performing a posterior belief assessment is in considering all possible alternatives of prior and likelihood to the original set of judgements J_0 that you would be unwilling to rule out based on your current understanding. As illustrated by our application to the ocean model, in complex models the ways in which the model can be plausibly (in your view) changed can quickly grow and be difficult to think about and to then group into co-exchangeable classes. There are many statistical models being developed that are far more complex than our ocean model example (but that can be readily sampled from using MCMC), so this task could seem far more daunting than in our case in some applications. However, in subjective Bayes there is no such thing as a free lunch and, making meaningful belief statements updated properly by available data, necessarily involves careful thought so that their meaning is well understood following the analysis.

The other main challenge in performing a posterior belief assessment involves obtaining the quantities required to compute $\mathrm{E}_{\mathcal{G}}[y]$, namely $\mathrm{E}[y]$, $\mathrm{Cov}[y, \mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{E}[\mathcal{G}]$. If these could not be elicited directly using partial prior specification methods discussed in chapter 2 of Goldstein and Wooff (2007), a combination of expert judgement (for quantities only involving y), and a sampling scheme we outlined in Section 4.2 could be used. The sampling method required many many more Bayesian computations to be performed (perhaps in more approximate form, e.g. MCMC with fewer samples), and may be infeasible if access to distributed computing is unavailable or in particularly complex problems. In particular, the number of Bayesian calculations required for a posterior belief assessment using our sampling algorithm will require a certain amount of automation in order to be feasible. Distributed computing programs such as condor make this feasible, as highlighted in our application. Our application was reasonably complex, yet we managed to perform 13000×72 separate Bayesian analyses based on MCMC using the automation provided by the condor program and a 1000 core cluster at Durham University in only 1 day. Developing tools to allow elicitation of these quantities directly, or developing alternative methods of deriving them through further, less computationally burdensome, calculations could be an important avenue of further investigation in this area.

References

Andiranakis, I. and Challenor, P. G. (2012), The effect of the nugget on Gaussian process emulators of computer models. *Computational Statistics and Data Analysis*, 56: 4215–4228. MR2957866. doi: http://dx.doi.org/10.1016/j.csda.2012.04.020. 882, 898

Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J., and Walsh, D. (2007), Computer model validation with functional output, *The Annals of Statistics*, 35, 1874–1906. MR2363956. doi: http://dx.doi.org/10.1214/009053607000000163. 898

Berger, J. O. (1994), An overview of robust Bayesian analysis, *Test*, 3(1), 5–124. MR1293110. doi: http://dx.doi.org/10.1007/BF02562676. 879

Berger, J. O. (2006), The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3), 385–402. MR2221271. 878

Craig, P. S., Goldstein, M., Rougier, J. C., and Seheult, A. H. (2001), Bayesian forecasting for complex systems using computer simulators, *Journal of the American Statistical Association*, 96, 717–729. MR1946437. doi: http://dx.doi.org/10.1198/016214501753168370. 898

Cumming, J. A. and Goldstein, M. (2009), Small sample designs for complex high-dimensional models based on fast approximations, *Technometrics*, 51, 377–388. MR2756474. doi: http://dx.doi.org/10.1198/TECH.2009.08015. 898

de Finetti, B. (1974), *Theory of Probability, Volume 1*, Wiley, New York. 884, 893

de Finetti, B. (1975), *Theory of Probability, Volume 2*, Wiley, New York. 884

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis, second edition*, Chapman and Hall/CRC Texts in Statistical Science. MR2027492. 879

Goldstein, M. (1997). Prior inferences for posterior judgements. In: *Structures and Norms in Science*, M. C. D. Chiara et al., eds., Kluwer, 55–71. MR1819330. 885

Goldstein, M. (2006), Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1(3), 403–420. MR2221272. 878

Goldstein, M. (2011). External Bayesian analysis for computer simulators. In: *Bayesian Statistics 9*. Bernado, J. M. et al., eds., Oxford University Press. MR3204007. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0007. 886

Goldstein, M. (2012). Observables and models: exchangeability and the inductive argument. In: *Bayesian Theory and Applications*, Damien, P. et al., eds., Clarendon Press, Oxford, 3–18. MR3221155. doi: http://dx.doi.org/10.1093/acprof:oso/9780199695607.003.0001. 893

Goldstein, M. and Wooff, D. (2007), *Bayes Linear Statistics Theory and Methods*, John Wiley and Sons Ltd. MR2335584. doi: http://dx.doi.org/10.1002/9780470065662. 879, 885, 888, 889, 891, 894, 903, 905

Haylock, R. and O'Hagan, A. (1996), "On inference for outputs of computationally expensive algorithms with uncertainty on the inputs." In: *Bayesian Statistics 5*, Bernado, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., eds., Oxford University Press, 629–637. MR1425432. 898, 899

Higdon, D., Nakhleh, C., Gattiker, J., and Williams, B. (2008), "A Bayesian calibration approach to the thermal problem", *Computer Methods in Applied Mechanics and Engineering,* 197, 2431–2441. 882

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, V. T. (1999), Bayesian model averaging: a tutorial, *Statistical Science*, 14(4), 382–401. MR1765176. doi: http://dx.doi.org/10.1214/ss/1009212519. 880, 889

Ingleby, B. and Huddleston, M. (2007), Quality control of ocean temperature and salinity profiles – Historical and real time data, *Journal of Marine Systems*, 65 158–175. 896

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. A. (2011), Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology, *The Annals of Applied Statistics,* 5(4), 2470–2492. MR2907123. doi: http://dx.doi.org/10.1214/11-AOAS489. 898

Kennedy, M. C. and O'Hagan, A. (2001), Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 63: 425–464. MR1858398. doi: http://dx.doi.org/10.1111/1467-9868.00294. 880, 881, 882, 898, 900

Lee, L., Pringle, K., Reddington, C., Mann, G., Stier, P., Spracklen, D., Pierce, J., and Carslaw, K. (2013), The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei, *Atmospheric Chemistry and Physics*, 13.17, 8879–8914. 898

Liang, F., Liu, C., and Carroll, R. J. (2010), *Advanced Markov Chain Monte Carlo Methods*, John Wiley and Sons Ltd., Chichester, UK. MR2828488. doi: http://dx.doi.org/10.1002/9780470669723. 878

Lindley, D. V. (2000), The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician),* 49(3), 293–337. 877

Madec, G. (2008), NEMO ocean engine. *Note du Pole de Modélisation.* Institut Pierre-Simon Laplace (IPSL), France, No. 27, ISSN No. 1288-1619. 896

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014), A web-based tool for eliciting probability distributions from experts. *Environmental Modelling and Software*, 52: 1–4, ISSN 1364-8152, http://dx.doi.org/10.1016/j.envsoft.2013.10.010. 878, 899

Oakley, J. E. and O'Hagan, A. (2010). *SHELF: the Sheffield Elicitation Framework (Version 2.0)*, School of Mathematics and Statistics, University of Sheffield (2010), www.tonyohagan.co.uk/shelf 878

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), Design and analysis of computer experiments, *Statistical Science*, 4, 409–435. MR1041765. 898

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag New York. MR2160708. doi: http://dx.doi.org/10.1007/978-1-4757-3799-8. 898

Savage, L. J. (1977) The shifting foundations of statistics. In: *Logic, Laws and Life: Some Philosophical Complications*, R.G. Colodny, ed., Pittsburgh University Press, Pittsburg, 3–18. 877

Sexton, D. M. H., Murphy, J. M., and Collins, M. (2011), Multivariate probabilistic projections using imperfect climate models part 1: outline of methodology, *Climate Dynamics*, doi: http://dx.doi.org/10.1007/s00382-011-1208-9. 898

Vernon, I., Goldstein, M., and Bower, R. G. (2010), "Galaxy formation: a Bayesian uncertainty analysis," *Bayesian Analysis,* 5(4), 619–846, with discussion. MR2740148. doi: http://dx.doi.org/10.1214/10-BA524. 898

Williamson, D. and Blaker, A. T. (2014), Evolving Bayesian emulators for structurally chaotic time series with application to large climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1) 1–28. MR3283898. doi: http://dx.doi.org/10.1137/120900915. 882

Williamson, D., Goldstein, M., and Blaker, A. (2012), "Fast linked analyses for scenario based hierarchies," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(5), 665–692. MR2993504. doi: http://dx.doi.org/10.1111/j.1467-9876.2012.01042.x. 898

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P. Jackson, L., and Yamazaki, K. (2013), History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics* 41: 1703–1729. doi: http://dx.doi.org/10.1007/s00382-013-1896-4. 898

Williamson, D. (2015). Exploratory ensemble designs for environmental models using k-extended Latin hypercubes *Environmetrics*, 26(4) 268–283. 882, 896

Williamson, D., Blaker, A. T., Hampton, C., and Salter, J. (2014). Identifying and removing structural biases in climate models with history matching. *Climate Dynamics*, Online First, doi: http://dx.doi.org/10.1007/s00382-014-2378-z. 898

Williamson, D., Blaker, A. T., and Sinha, B. (2015) Statistical ocean model tuning and parametric uncertainty quantification with NEMO, *In preparation*. 899

Acknowledgments

Danny Williamson is funded by an EPSRC fellowship grant number EP/K019112/1. We wish to thank an associate editor and two referees for their thoughtful and detailed comments regarding the paper. Their suggestions have helped to improve it a great deal. We also thank Adam Blaker and Bablu Sinha for running the NEMO ocean model to generate the ensemble that appears in our application.