Bayesian Analysis (2022) **TBA**, Number TBA, pp. 1–33

A Bayesian Computer Model Analysis of Robust Bayesian Analyses

Ian Vernon* and John Paul Gosling†

Abstract. We harness the power of Bayesian emulation techniques, designed to aid the analysis of complex computer models, to examine the structure of complex Bayesian analyses themselves. These techniques facilitate robust Bayesian analyses and/or sensitivity analyses of complex problems, and hence allow global exploration of the impacts of choices made in both the likelihood and prior specification. We show how previously intractable problems in robustness studies can be overcome using emulation techniques, and how these methods allow other scientists to quickly extract approximations to posterior results corresponding to their own particular subjective specification. The utility and flexibility of our method is demonstrated on a reanalysis of a real application where Bayesian methods were employed to capture beliefs about river flow. We discuss the obvious extensions and directions of future research that such an approach opens up.

Keywords: emulation, Gaussian process, sensitivity analysis.

1 Introduction

Bayesian methodology is now widely employed across many scientific areas (for example, over 490 articles have been published in Nature containing the word "Bayesian"; Springer Nature, 2022). The uptake of Bayesian methods is due both to progress in Bayesian theory and to advances in computing power combined with the development of powerful numerical algorithms, such as Markov Chain Monte Carlo (MCMC). However, many Bayesian analyses of real world problems are both complex and computationally time-consuming. They often involve complex hierarchical models that require large numbers of structural and distributional assumptions both in the likelihood and prior (along with other choices covering the numerical implementation). Due to the long run times and the need to tune such algorithms, it is common for little or no rigorous sensitivity analysis to be performed, therefore it is often unclear as to what extent the Bayesian posterior and the subsequent decisions it informs have been affected by these numerous assumptions. For any serious scientific analysis, a solid understanding of the inferential process and its response to changes in the underlying judgements and assumptions is absolutely vital. Any Bayesian analyses that cannot do this is of limited use and, we would assert, has questionable worth to the scientific community.

Much work has been done to address the issues of robustness and sensitivity analysis of Bayesian analyses, with many elegant results derived (see for example Box and

[*] Department of Mathematical Sciences, Durham University, Stockton Road, Durham, DH1 3LE, UK, i.r.vernon@durham.ac.uk

[†] Department of Mathematical Sciences, Durham University, Stockton Road, Durham, DH1 3LE, UK, john-paul.gosling@durham.ac.uk

https://doi.org/10.1214/22-BA1340

Tiao, 1962; Berger, 1994; Berger et al., 2000; Roos et al., 2015). However, progress in this area has greatly slowed over the past fifteen years due in part to the intractability of analysing even fairly basic Bayesian models. In particular, although aspects of prior sensitivity were explored (see e.g. Berger, 1994; Moreno, 2000; Fan and Berger, 2000) and loss sensitivity (Dey and Micheas, 2000), perturbations to the likelihood proved far more challenging to deal with analytically (Shyamalkumar, 2000). Two broad robust Bayesian strategies can be distinguished, the first of these being the global approach, whereby whole classes of priors and/or likelihoods are considered, and their effects on the posterior analysed. While there was much early success in this direction (see for example Berger and Sellke, 1987; Berger, 1994; Moreno, 2000), many of these results relied upon appeals to monotonicity arguments which were of great use in lower dimensional cases, but less easy to apply in more complex, higher dimensional models. Even defining sensible prior or likelihood classes to investigate in high dimension, while avoiding vacuous results, becomes problematic (Insua and Ruggeri, 2000). See also Kallioinen et al. (2021) for a power-scaling approach. Increasing attention was also directed at a second strategy, that of the local sensitivity approach, whereby the derivatives of posterior features of interest with respect to perturbations of various forms are analysed often using differential calculus (Gustafson and Wasserman, 1995; Gustafson, 2000; Perez et al., 2005; Zhu et al., 2007; Muller, 2012; Roos and Held, 2011; Giordano et al., 2018). While far more tractable, the local approach has obvious weaknesses, in that its results may be strongly sensitive to the original prior and likelihood specification. For many complex Bayesian models, for which the posterior features may be highly non-linear functions of the perturbations, such local approaches will be clearly inadequate.

Despite the efforts of the robust community, it must be conceded that the huge advances in MCMC and comparable numerical methods, which allow the use of more and more complex Bayesian models, have left robust Bayesian analysis techniques far behind (Watson and Holmes, 2016; Robert and Rousseau, 2016). As complex Bayesian models along with MCMC algorithms are now widely used in areas of real world importance, and as our Bayesian community will be judged upon the results of these algorithms, the need for powerful, general robust methods applicable to a wide class of perturbations is increasingly urgent. This article suggests a framework for the solution to this problem. We propose to treat a complex and computationally demanding Bayesian analysis as an expensive computer model. We utilise Bayesian emulation technology developed for complex computer models (as described in O'Hagan, 2006, for example) to explore the structure of the Bayesian analysis itself, and, specifically, its response to various changes in both the prior and likelihood specification. This allows for a more general sensitivity and robustness analysis that would be otherwise unattainable, because we do not require analytic solutions. This methodology is very flexible, provides both local and global results, is straightforward to implement in its basic form using currently available emulation software packages, and can deal with a wide class of statistical analyses.

In more detail, a typical Bayesian analysis involves many judgements and assumptions, both in relation to modelling choices that feed into the likelihood and in terms of the representation of prior beliefs. Often, pragmatism leads to assumptions being made that are based either entirely or in part on mathematical convenience. For example, conjugate analyses whereby convenient mathematical forms are chosen in both the likelihood and prior. Aside from modelling choices, expressing judgements in probabilistic form can be time consuming and difficult, so in many cases tractable representations followed by simple assessments are made that only approximately represent the beliefs of the expert. At the other extreme, so-called objective priors are used either due to their reported properties, or because any relevant subjective beliefs are thought to be too weak to alter the analysis to any noticeable degree. All of the above compromises are defensible only if it can be shown that the posterior attributes of interest are relatively insensitive to small changes in the prior and modelling specifications. Our approach is to explore the concerns regarding the specific choices and assumptions used to form the prior and modelling specifications by embedding the current Bayesian analysis within a larger structure, constructed by parameterising the major set of choices made, following the robust Bayesian paradigm. This larger structure is then subjected to Bayesian computer model techniques (which, as discussed below, implies a second layer of Bayesian analysis is being used). While not all choices can be parameterised directly, as we will discuss, often the major sources of concern can be addressed in this way. We note that Peruggia et al. (2004) employed GPs to explore prior robustness issues, but were limited to a small number of prior quantities. Here, we argue for a general analysis which examines both prior and (the more analytically challenging) likelihood quantities, and their various interactions, simultaneously in a comprehensive global approach.

Our approach also addresses another major concern: that of multiple subject area experts, who each may possess different beliefs regarding the prior and likelihood structures. Even when a thorough Bayesian analysis, possibly using MCMC, is performed and published, its results are usually based on the judgements of a single expert (or small group of experts). It is therefore difficult for other experts in the area to know how to interpret these results: what they really require is for the MCMC to be rerun with their beliefs inserted instead. Therefore, at the very least, the statistician should facilitate the analysis of a class of prior or likelihood statements, approximately representing the differing views held across the relevant scientific community (Aczel et al., 2020). Unfortunately this is not provided in the vast majority of Bayesian analyses, albeit due to understandable constraints on time and computational resources. However, our analysis will enable experts to quickly extract approximations to their posterior results corresponding to their own specification, along with associated uncertainty statements. Importantly, this is straightforward to implement, and only requires the off-line running of existing MCMC code with minor adaptations, in embarrassingly parallel fashion, leading to minimal increase in wall-clock time when cluster or cloud resources are available. This approach therefore provides what many scientific fields require: complex Bayesian analyses that are simultaneously applicable to a range of scientific specifications.

The article is organised as follows. In Section 2 we recast the problem of a robust Bayesian analysis into that of a complex computer model, describe computer model emulation methodology, and then apply it to an example Bayesian model. In Section 3 the utility and flexibility of our method is demonstrated on a reanalysis of a real application concerning river flow. We discuss the various choices one faces in this kind of analysis, and outline several areas of future research in Section 4, before concluding in Section 5. Code for reproducing the example Bayesian model featured throughout

Section 2 and the supplementary material (Vernon and Gosling, 2022), is provided at https://github.com/ivernon/BARBA.git.

2 Bayesian analysis as a complex computer model

Our set-up is similar in structure to that of a robust Bayesian analysis; however, we utilise a computer model representation and notation (see for example Craig et al., 1997; Kennedy and O'Hagan, 2001; Higdon et al., 2004; Vernon et al., 2010a). Let us assume that interest lies in a vector of random quantities θ, beliefs about which will be updated in the light of a vector of data z. The prior $\pi(\theta|x_p)$ and likelihood $l(z|\theta, x_l)$ are both conditioned on some specific list of choices and modelling assumptions represented by parameters x_p and x_l respectively, an example of which would be hyper-parameters that have been kept constant. We wish to explore features of interest of the posterior $\pi(\theta|z, x_p, x_l)$ such as the mean, variance, quantiles, etc. chosen due to their relevance to the downstream application or decision process. We map the posterior to this vector of attributes using the functional $g(.)$ and, hence, define the overall function $f(x)$ as:

$$f(x) \;\; = \;\; f(x_p, x_l) \;\; = \;\; g(\pi(\theta|z, x_p, x_l)) \tag{1}$$

where $x = (x_p, x_l)$ is the combined vector of inputs that parameterise the specific choices and assumptions made in the prior and likelihood specifications, and $f(x)$ is the vector of all posterior features and summaries of interest, where the dependence on the data z is now implicit. Note that it would also be simple to extend equation (1) to include a loss function, corresponding inputs to the loss, and various decision end points (Oakley, 2009). An example of $f(x)$ that we use in Section 2.2, where the posterior mean and standard deviation are of primary interest is:

$$f(x) \; = \; (\mathrm{E}[\theta|z, x], \mathrm{SD}[\theta|z, x])\,. \tag{2}$$

For most Bayesian analyses, in order to evaluate the posterior, we require a possibly expensive sampling algorithm such as MCMC, which may take hours, days or even weeks for one evaluation for a particular choice of inputs x. Hence, we can view the implementation of the Bayesian analysis as an expensive computer model $f(x)$, that maps a possibly high dimensional input vector x to a vector of outputs f of primary interest to the modeller. Note that we would be free to view the MCMC algorithm itself as a stochastic computer model, in which case we could add any algorithm inputs x_{MCMC} such as parameters governing the adaptive regime, burn-in and so on to the input vector x, and include additional diagnostic outputs into the vector f such as the MCMC acceptance rates. However, we leave such complications to future work, as here, we are primarily interested in the key features $f(x)$ of the underlying Bayesian analysis itself, which the MCMC output only approximates. The precise representation of the link between the MCMC output and $f(x)$ will be given in Section 2.1.

We then seek to explore the behaviour of the posterior features of interest $f(x)$ as a function of the inputs x across a wide class of Bayesian analyses defined as

$$\mathcal{F} \; = \; \{f(x) : x \in \mathcal{X}\} \tag{3}$$

where $\mathcal{X}$ governs the extent of our robust-Bayesian analysis and allows us to explore simultaneous changes in the prior and likelihood specifications. Note that in general, for a high dimensional and large enough $\mathcal{X}$, we would expect both the location and shape of the posterior $\pi(\theta|z, x)$ to vary substantially over $\mathcal{X}$, and hence that standard techniques based around re-sampling an individual MCMC sample (see for example Smith and Gelfand, 1992; Geweke, 1999), or importance sampling (see for example Geyer, 1994; Fan and Berger, 2000; Sinharay and Stern, 2002), may not be effective. See Sup. Mat. section 2.3, for further discussion and an illustrative comparison.

We envisage that the need to explore a class of Bayesian analyses may arise for several reasons: for example, we may wish to perform a global robust Bayesian analysis over $\mathcal{X}$ due to a possibly imprecise specification or to perform a local sensitivity analysis. Alternatively, we may be dealing with a collection of experts whose opinions on the prior and likelihood differ, but which are all contained within $\mathcal{X}$. Therefore, we depart somewhat from the goal of a typical robust analysis in that we are primarily interested in the entire behaviour of $f(x)$ over the set $\mathcal{X}$, and not just in the extrema of $f(x)$. This is because we want our results to be applicable for any user that has a precise or imprecise specification contained within $\mathcal{X}$, and because we may also wish to understand and identify any sensitive regions where $f(x)$ rapidly changes as a function of x. Unlike in many computer model analyses, we therefore do not view x as being uncertain: if this was the case we would simply build an additional layer of prior structure over x into our Bayesian hierarchical model (which, incidentally our techniques would facilitate). Instead, we seek to efficiently represent, using an emulator, the behaviour of $f(x)$ for any value of $x \in \mathcal{X}$. If an expert subsequently came with their own specification x_e, they would instantly be able to read off the likely values of the posterior features of interest $f(x_e)$ corresponding to their own particular beliefs. Additionally, the results of our analysis should provide approximate answers to any local robustness, global robustness or sensitivity analysis question regarding $f(x)$, critically, with an attached statement of uncertainty. The emulator structure that incorporates this uncertainty can also guide future evaluations of the sampling algorithm designed to resolve key uncertainties of most interest to the expert(s). As we attempt to represent a large class of inputs and outputs, our approach is more general than a perfunctory robust Bayesian analysis, and should be widely applicable. We now go on to describe the emulation process, and how to adapt it for application to the analysis of Bayesian analyses.

2.1 Computer model emulation

Here we give a brief overview of computer model emulation: see the Sup. Mat. for a more detailed introduction, including a list of suitable emulation packages. Gaussian process emulation is a powerful technique for modelling and subsequently analysing expensive computer models that may possess high dimensional input and output spaces (see for example Craig et al., 1997; Kennedy and O'Hagan, 2001; Heitmann et al., 2009; Vernon et al., 2010a,b; Andrianakis et al., 2015; Gu and Berger, 2016; Edwards et al., 2021). The computer model is viewed as an expensive function that maps a vector of inputs x to a vector of outputs $f(x)$. Beliefs about the value of the uncertain function $f(x)$ at an untried input x are represented by a Gaussian process prior over $f(x)$, also termed an emulator,

$$f(.)|m(.), c(.,.) \quad \sim \quad GP(m(.), c(.,.)), \tag{4}$$

with a mean function $m(.)$ capturing global behaviour, and a covariance function $c(.,.)$ representing the local smoothness of $f(x)$, taking, for example, Gaussian form

$$c(x, x') \quad = \quad \sigma^2_{em} \exp\{-||x - x'||^2/\theta^2_{em}\}, \tag{5}$$

where σ^2_{em} and θ_{em} are emulation parameters that need to be specified. Other forms for $c(.,.)$ including the much used Matern function, are of course available (Santner et al., 2003). A design of runs is performed at n input locations $x_D = \{x_D^{(1)}, \ldots, x_D^{(n)}\}$ over the d-dimensional input space $\mathcal{X}$ giving a vector of outputs $f(x_D)$. The precise number n of runs required will depend upon the nature of the model and the desired emulator accuracy, but a rough guide is to use at least $10d$ runs, as argued by Loeppky et al. (2009), in a space-filling design, as discussed in Santner et al. (2003). The emulator is then updated by $f(x_D)$, and the posterior mean and covariance function obtained

$$f(.)|f(x_D), m(.), c(.,.) \quad \sim \quad GP(m^*(.), c^*(.,.)), \tag{6}$$

$$m^*(x) \quad = \quad m(x) + \text{Cov}\,[f(x), f(x_D)]\,\text{Var}[f(x_D)]^{-1}(f(x_D) - \text{E}[f(x_D)]), \tag{7}$$

$$c^*(x, x') \quad = \quad c(x, x') - \text{Cov}\,[f(x), f(x_D)]\,\text{Var}[f(x_D)]^{-1}\text{Cov}\,[f(x_D), f(x')]\,. \tag{8}$$

Evaluation of the emulator, in terms of its mean and variance, for different values of x, is usually several orders of magnitude faster that the original computer model, hence the behaviour of $f(x)$ can be investigated far more thoroughly, and sensitivity analysis, history matching, calibration and many other powerful techniques can be performed (Oakley and O'Hagan, 2004; Kennedy and O'Hagan, 2001; Vernon et al., 2010a,b). More advanced forms of the emulator are of course possible, possessing a more structured mean function and exploiting the concept of active inputs, which helps combat the problems associated with high input dimension (Vernon et al., 2010a,b). Another useful feature of Gaussian process emulation is its representation of derivatives. If the computer model function $f(x)$ is a Gaussian process, then the partial derivatives $\partial f(x)/\partial x_i$ also form Gaussian processes, with covariance function naturally constructed by taking the partial derivatives of $c(.,.)$ (O'Hagan, 1992), a feature that we will exploit. Although Gaussian processes are perhaps the most popular tool for emulation, several other approaches are available e.g. the related Bayes linear version (Craig et al., 1997; Vernon et al., 2010a), BART methods (Chipman et al., 2012), BMARS (Francom et al., 2018), Neural Networks (Grzeszczuk et al., 1998) and dynamic models (Liu and West, 2009).

Various diagnostics are available to check emulator performance(Bastos and O'Hagan, 2008). Once a satisfactory emulator has been constructed, Variance-based sensitivity indices (Saltelli et al., 2000) can be calculated efficiently using the probabilistic sensitivity analysis techniques described in Oakley and O'Hagan (2004). The sensitivity indices can be used to give an indication of which model inputs are responsible for most variation in the model outputs (given the range of plausible values for the inputs): the main-effect indices give the proportion of variance in the output explained by a input acting on its own and the total-effect indices give the proportion of variance in the output explained by a input on its own and in conjunction with other inputs. See also Francom et al. (2018) where Bayesian MARS emulators are used, which helpfully allow for analytic sensitivity analysis calculations.

Adapting emulation for application to a Bayesian analysis

All of the above emulation methodology can, with slight modification, be applied to the outputs of an MCMC algorithm as part of a robust Bayesian analysis, as represented by $f(x)$ with $x \in \mathcal{X}$, or indeed to any statistical analysis that is expensive to perform and for which one requires a sensitivity analysis.

We would start by designing a space filling batch of n runs $x_D = \{x_D^{(1)}, \ldots, x_D^{(n)}\}$ over the d-dimensional input space $\mathcal{X}$. The MCMC algorithm would then be run at each of the design points, and the usual convergence tests and examination of mixing plots would be performed. Note that our framework can of course incorporate information from alternative MCMC algorithms, as we discuss in Section 4.2, however convergence issues may favour the approach described here. Due to the large number of burn in steps required for MCMC convergence, a suitable design would most likely favour a smaller number of design points with a large number of posterior samples drawn at each point: a classic computer model set up. Here, we use space filling designs (see for example, Morris and Mitchell, 1995), with large numbers of posterior samples, and leave a more detailed treatment of such design questions to future work.

An important difference from the standard deterministic computer model emulation setup is that, as the MCMC algorithm only returns draws from the posterior, it should be viewed as a stochastic computer model, and hence allowance made for the fact that we only see, for example, sample means and sample variances and not the true posterior values. There are many approaches to the emulation of stochastic computer models of varying complexity (see for example Johnson et al., 2011; Andrianakis et al., 2017; Vernon and Goldstein, 2022). Here, we generate large MCMC samples and treat the resulting low level of stochasticity via a simple nugget representation. Although simple, we make the connection between the true posterior quantities and their MCMC sample counterparts explicit to facilitate a later discussion of the partial derivatives of $f(x)$, of use in a local sensitivity analysis. Representing the Bayesian posterior features of interest as $f(x)$ and the corresponding sample quantities obtained from the MCMC algorithm as $f^{(s)}(x)$, we model the link between the two for output i as:

$$f_i^{(s)}(x) \;=\; f_i(x) + \eta_i(x) \tag{9}$$

where $\eta_i(x)$ is an uncorrelated nugget term possessing zero mean and constant variance across the input space, usually estimated from the MCMC run data (Andrianakis et al., 2015). Note that as the effective sample size of the MCMC runs will be large, the variance of $\eta_i(x)$ will be far smaller than other uncertainties, and more detailed modelling will be in many cases unwarranted.

We may believe that $f_i(x)$ is smooth and, hence, choose an appropriate correlation structure for it, given say by equation (5). It follows that the correlation function for the MCMC output $f_i^{(s)}(x)$ becomes

$$\mathrm{Cov}\left[f_i^{(s)}(x), f_i^{(s)}(x')\right] = c^{(s)}(x, x') \;=\; \sigma_{em}^2\left[(1-\delta_{em})\exp\{-||x-x'||^2/\theta_{em}^2\} + \delta_{em}\delta_{x,x'}\right] \tag{10}$$

where $\delta_{x,x'} = 1$ when $x = x'$ and 0 otherwise, δ_{em} controls the influence of the nugget variance, and σ^2_{em} now represents the prior variance of $f_i^{(s)}(x)$. The covariance between $f_i(x)$ and $f_i^{(s)}(x)$ is now

$$\mathrm{Cov}\left[f_i(x), f_i^{(s)}(x')\right] \;=\; \sigma^2_{em}(1-\delta_{em})\exp\{-||x-x'||^2/\theta^2_{em}\} \tag{11}$$

We can construct an emulator for $f_i(x)$ as before, using the expressions for the posterior mean and correlation given by equations (7) and (8), but now we replace all occurrences of $f(x_D)$ by $f^{(s)}(x_D)$ in equations (7) and (8), and use equations (10) and (11) to evaluate the altered covariance terms.

Another benefit of this construction, where we have implicitly included the smoothness of $f(x)$ (noting that $f^{(s)}(x)$ is of course not smooth), is that we can also construct emulators for the partial derivatives $\partial f(x)/\partial x_j$ for minimal extra computational cost. These follow the same principals, but with the correlation between the derivatives $\partial f(x)/\partial x_j$ and the MCMC outputs $f^{(s)}(x)$ now given for output i by:

$$\mathrm{Cov}\left[\frac{\partial f_i(x)}{\partial x_j}, f_i^{(s)}(x')\right] \;=\; -\frac{2}{\theta^2}\sigma^2_{em}(1-\delta_{em})(x_j - x'_j)\exp\{-||x-x'||^2/\theta^2_{em}\} \tag{12}$$

which is obtained by partially differentiating equation (11) (O'Hagan, 1992). The derivative emulators are evaluated using equations (7) and (8) as before, but now with $f(x)$ replaced by $\partial f(x)/\partial x_j$.

Once the emulators have been constructed, they can be used to explore the behaviour and both the local and global sensitivity of the outputs of the Bayesian analysis $f(x)$ to the decisions made, as represented by the inputs x. It is worth noting that there are now *two* distinct Bayesian processes at work here (as highlighted by the double use of the word "Bayesian" in the article title): the updating of our prior beliefs about the function $f(x)$ via the Gaussian process emulator structure, and the update of the original Bayesian problem that occurs at every single point in the $\mathcal{X}$ space, that is the update of θ by z (conditioned on x). Using the emulators, questions of robustness can be addressed, and various graphical methods can be employed to explore these, which we develop in Sections 2.2 and 3.2. Many other useful computer model techniques still have important analogies in this setting e.g. history matching, model discrepancy and calibration, and we discuss their uses in Section 4. We now go on in the next section to demonstrate our techniques on an example Bayesian model.

2.2 Example Bayesian model

We introduce a synthetic example of a Bayesian analysis to demonstrate the proposed methodology. Despite the simplicity of this model, it exhibits some interesting features in terms of the response of the posterior to the prior and likelihood specification, that highlight the utility of our approach. We investigate a case in which we imagine there is reasonable disagreement between experts over both prior and likelihood specifications.

Scalar data $z_i > 0$ with $i = 1 \ldots 10$ are observed, and we imagine that a Bayesian analysis has initially been performed with the following conjugate specification. The data, given in appendix A of the supplementary material, are assumed to be independent and identically distributed with likelihood given by

$$\begin{aligned} z_i|\theta &\sim \text{Exp}(\theta), \qquad i = 1,\dots,10 \qquad (13)\\ \Rightarrow \pi(z_i|\theta) &= \theta\, e^{-\theta z_i}, \qquad z_i \geq 0 \qquad (14)\end{aligned}$$

parameterised in terms of the rate parameter θ, which has corresponding prior

$$\theta|\mu,\nu^2 \;\sim\; \text{Ga}(\mu,\nu^2) \qquad (15)$$

where $\text{Ga}(\mu,\nu^2)$ denotes a gamma distribution that has been parameterised in terms of its mean and variance, μ and ν^2 respectively. Initially, the prior hyperparameters were judged to be $\mu_0 = 5$ and $\nu_0 = 1$.

Given data, this Bayesian analysis would be easy to implement given that the prior distribution is conjugate. We imagine that there is however concern amongst the experts about the data generating process, specifically with the tails of the likelihood and its behaviour close to $z_i = 0$. We explore these concerns by contaminating the likelihood with a half-normal component $z_i|\theta \sim HN(\theta)$, where the impact of the contamination is controlled by a mixing parameter $\epsilon \in [0,1]$. When $\epsilon = 1$ the likelihood is purely half-normal so that

$$\begin{aligned} z_i|\theta,\epsilon{=}1 &\sim HN(\theta), \qquad i = 1,\dots,10 \qquad (16)\\ \Rightarrow \pi(z_i|\theta,\epsilon{=}1) &= \frac{2}{\pi}\,\theta\, e^{-\theta^2 z_i^2/\pi}, \qquad z_i \geq 0 \qquad (17)\end{aligned}$$

where we have parameterised the half-normal distribution in terms of its inverse mean θ, such that $\text{E}[z_i|\theta,\epsilon{=}1] = 1/\theta$, in direct agreement with the definition of θ in the uncontaminated exponential likelihood of equations (13) and (14). The full contaminated likelihood for $z = (z_1,\dots,z_{10})$, conditioned on the contamination parameter ϵ, can now be written as

$$\pi(z|\theta,\epsilon) \;=\; \prod_{i=1}^{10}\left((1-\epsilon)\,\theta\, e^{-\theta z_i} + \epsilon\,\frac{2}{\pi}\,\theta\, e^{-\theta^2 z_i^2/\pi}\right). \qquad (18)$$

where we have ensured that the property $\text{E}[z_i|\theta,\epsilon] = 1/\theta$ still holds for any z_i and now any ϵ, consistent with the original specification. Figure 1 (left panel) shows $\pi(z|\theta,\epsilon)$ as a function of θ for various levels of ϵ. The contamination parameter ϵ represents, and is used to investigate, the experts' disagreement over the structure of the likelihood, and returns it to the pure exponential form and hence to conjugacy as $\epsilon \to 0$. The experts are still satisfied with a gamma prior and agree with the prior mean $\mu_0 = 5$, but not with the prior variance $\nu^2 = 1$ for which there is a range of alternative opinions:

$$\theta|\mu_0,\nu^2 \;\sim\; \text{Ga}(\mu_0,\nu^2), \qquad 0.3 < \nu < 2; \qquad (19)$$

hence, ν now parameterises differing levels of prior uncertainty.

The above description specifies a simple class of possible Bayesian analyses defined over a 2-dimensional space $\mathcal{X}$ where

$$\mathcal{X} \;\equiv\; \{x = (\nu,\epsilon) \;:\; \nu \in [0.3,2] \text{ and } \epsilon \in [0,1]\}, \qquad (20)$$

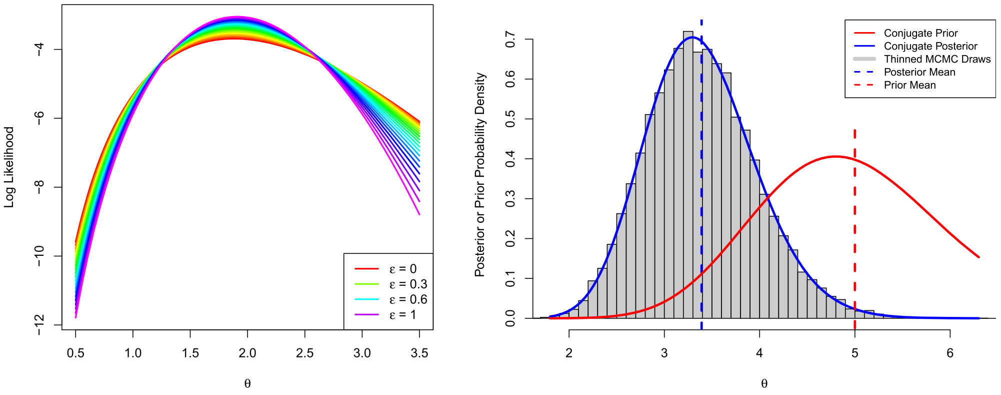

Figure 1: Left panel: the contaminated log-likelihood $\log(\pi(z|\theta, \epsilon))$, given by equation (18), as a function of θ, with the colours labelling differing values of the contamination parameter ϵ. Right panel: draws from the MCMC algorithm in the original conjugate case (when $\nu = 1$ and $\epsilon = 0$), showing the theoretical posterior density $\pi(\theta|z, \nu\!=\!1, \epsilon\!=\!0)$ in blue and the prior density $\pi(\theta|\nu\!=\!1)$ in red. The prior and posterior means are given as the vertical dashed lines in red and blue respectively, the later is the first output $f_1(x)$ to be emulated. This plot therefore represents the single point $x = (1, 0)$ in the space of possible Bayesian analyses denoted by $\mathcal{X}$.

Inputs x	Type of Input	Range	Outputs $f(x)$
ν	Prior standard deviation	$[0.3, 2]$	$\mathrm{E}[\theta|z, \nu, \epsilon]$
ϵ	Likelihood contamination	$[0, 1]$	$\mathrm{SD}[\theta|z, \nu, \epsilon]$

Table 1: The inputs x and outputs $f(x)$ of the example Bayesian model when represented as a computer model. The classes of inputs and outputs are also given along with the range of exploration of the inputs, defining the extent of the sensitivity analysis.

which is parameterised by the likelihood contamination and prior standard deviation parameters, ϵ and ν respectively, as summarised in Table 1. We now wish to explore the behaviour of attributes of the posterior $\pi(\theta|z, \nu, \epsilon)$ as a function of the inputs ν and ϵ, and to investigate the corresponding robustness of these attributes and, hence, of the original analysis. We choose here to examine the posterior mean and standard deviation as these are usually of primary interest, but our approach could be applied to any set of posterior attributes: see Sup. Mat. for an extension using quantiles. We define

$$f(x) \;=\; (\mathrm{E}[\theta|z, \nu, \epsilon], \mathrm{SD}[\theta|z, \nu, \epsilon]) \tag{21}$$

as the function to be explored, as also summarised in Table 1. Note that a perfunctory robust Bayesian analysis at this point may attempt to examine the range of possible values of the posterior attributes of interest, in this case the mean and standard deviation, that is achievable over $\mathcal{X}$. We wish to go further and to efficiently represent the posterior attributes for *any* choice of the inputs ν and ϵ. This allows any expert to be able to extract their own Bayesian posterior attributes directly from our results, either corresponding to a particular specification represented as a single point in $\mathcal{X}$, or to a range of possible specifications represented by a subset of $\mathcal{X}$.

As the specification is no longer in general conjugate, we construct a simple Metropolis-Hastings MCMC algorithm to allow evaluation of the posterior at any choice of inputs ν and ϵ. As such a sampling algorithm is in some sense expensive (or would be for larger, more realistic models), we view the Bayesian updating process and its MCMC implementation as an expensive computer model, represented as the function $f(x)$, and employ computer model methodology in order to emulate and analysis the behaviour of $f(x)$. As the parameter of interest θ is non-negative, a Metropolis-Hastings MCMC algorithm was employed with a folded normal proposal distribution,

$$\theta^*|\theta_{t-1}, \xi_\theta^2 \sim FN(\theta_{t-1}, \xi_\theta^2), \tag{22}$$

where the folded normal has location parameter θ_{t-1}, with the scale parameter fixed at $\xi_\theta^2 = 0.9$, which yielded reasonable acceptance rates between 0.30 and 0.59 for all evaluations of interest (Brooks et al., 2011). Note that the folded normal is still a symmetric proposal density, allowing for the usual simplification to the acceptance ratio. To avoid unnecessary complications in the description of our approach to this illustrative example, we minimised the MCMC sampling error and ensured convergence by running an excessively large number of steps. A total of 200000 steps were used, the initial condition $\theta = 0.5$ was chosen and a burn in of 100 steps assumed. An example of the posterior sample generated by the MCMC algorithm is given as the grey histogram in Figure 1 (right panel), for the initial conjugate case where $\epsilon = 0$ and $\nu = 1$. Also shown is the prior and true posterior distributions as the red and blue lines respectively. The prior mean μ_0 and posterior mean $\mathrm{E}[\theta|z, \nu{=}1, \epsilon{=}0]$ are given as the vertical dashed red and blue lines respectively. This figure therefore represents a single point in the class of Bayesian updates $\mathcal{X}$ that we wish to emulate over, the specific point being

$$f(x = (1,0)) = (\mathrm{E}[\theta|z, \nu = 1, \epsilon = 0], \mathrm{SD}[\theta|z, \nu = 1, \epsilon = 0]) \,.$$

For any point in $\mathcal{X}$ with $\epsilon > 0$, conjugacy is no longer true and MCMC becomes vital.

Emulating the Bayesian analysis

In order to construct an emulator for the function $f(x)$ over $\mathcal{X}$, we now run the MCMC algorithm at a set of 35 design points x_D over the two dimensional input space $\mathcal{X}$, using a lattice design (see Sup. Mat. Table 1). This choice of run number, slightly larger than the rough guide suggestion of $10d$, was made to ensure we obtained reasonably accurate emulators after a single batch of runs. We check the convergence, the mixing plots and the autocorrelation plots for each of the 35 MCMC chains. As θ here is one dimensional, and as we employed a very large number of steps, our MCMC algorithm was unsurprisingly found to perform adequately across the whole input space (we discuss alternate MCMC strategies in Section 4.2).

Figure 2 (left panel) shows the estimated posterior density functions for the 35 separate MCMC-based analyses performed across $\mathcal{X}$, which display a reasonable range of

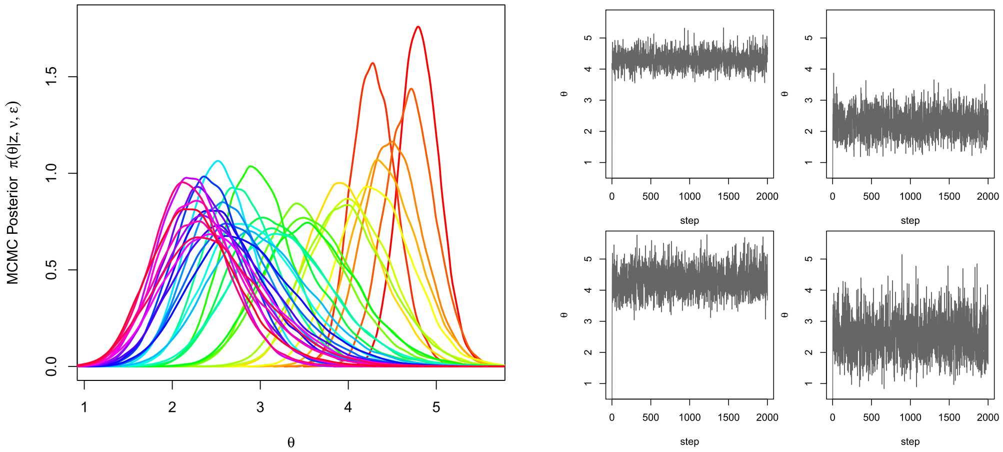

Figure 2: Left Panel: the estimated posterior density functions for the 35 MCMC runs performed across $\mathcal{X}$, coloured by their ν values, which display a reasonable range of posterior means and standard deviations. Right Panel: mixing plots for the four runs closest to the corners of the space $\mathcal{X}$, demonstrating excellent convergence and mixing, as expected.

posterior means and standard deviations. This implies that the different choices within the analysis, as represented by $\mathcal{X}$, will lead to substantial differences in the Bayesian posterior. Note that this would likely preclude alternative strategies based on re-weighting one posterior sample to estimate other posterior attributes across $\mathcal{X}$, strategies that would likely become even weaker for more complex problems: see Sup. Mat. section 2.3 for further discussion comparing our approach with importance sampling. Figure 2 (right panel) shows the mixing plots for the four runs closest to the corners of the space $\mathcal{X}$, and demonstrate convergence and excellent mixing, as expected.

With such checks in place, we are now free to emulate the function $f(x)$ over the input space $\mathcal{X}$ using the methodology described in Section 2. Specifically, we used a simple emulator construction sufficient for this example, with constant mean function $m(x) = m_0$, and covariance function $c(x, x')$ given by the Gaussian form of equation (5). We set the variance of the nugget equal to the mean of the MCMC sampling variance (a very small value), which was assumed constant across the input space. The inputs x were scaled to have range $[-1, 1]$ and a fixed correlation length of $\theta_{em} = 0.6$ was used for both, following the arguments in Vernon et al. (2010a,b) for choosing correlation lengths *a priori*. Finally, the emulator variance parameter σ^2_{em} was set equal to the variance of the 35 run outputs. See Sup. Mat. Table 1 for the design and MCMC output.

Figure 3 (left panel) shows the emulator expectation $\mathrm{E}[f_1(x)|f_1^{(s)}(x_D)]$, given by equation (7), for the posterior mean $f_1(x) = \mathrm{E}[\theta|z, x]$ as a function of the inputs $x = (\nu, \epsilon)$ (we suppress the implicit conditioning on $m(.)$ and $c(.,.)$ as given in equation (6) from here onward). The blue dot represents the original conjugate analysis where $x = (1, 0)$, the output of which is shown in Figure 1 (right panel). This plot in-

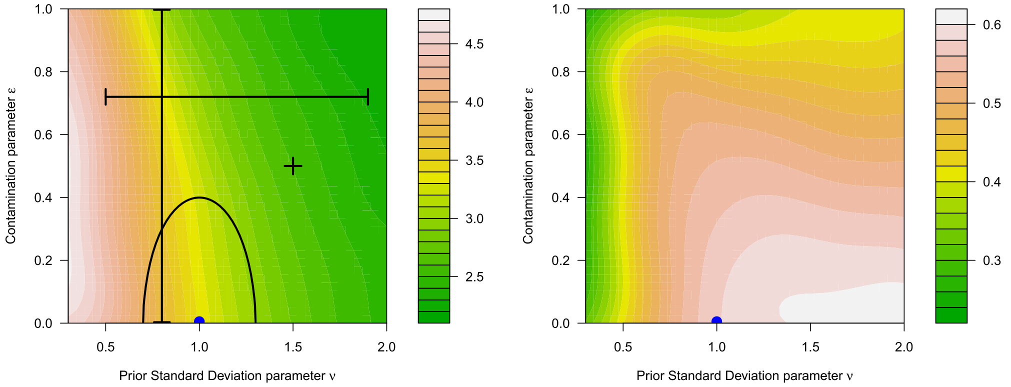

Figure 3: Left panel: the emulator expectation $\mathrm{E}[f_1(x)|f_1^{(s)}(x_D)]$ for the posterior mean $f_1(x) = \mathrm{E}[\theta|z,x]$ as a function of the inputs $x = (\nu, \epsilon)$, with the blue dot at the location of the original conjugate analysis, shown in Figure 1 (right panel). We see that the posterior mean is more sensitive to ν than to ϵ, and that the original conjugate analysis is relatively robust to small departures from conjugacy, of the given form. The black cross, vertical, horizontal and semi-elliptical lines represent the Cases 1-4 respectively, described in the text. Right panel: the emulator expectation $\mathrm{E}[f_2(x)|f_2^{(s)}(x_D)]$ for the posterior standard deviation $f_2(x) = \mathrm{SD}[\theta|z,x]$, as a function of the inputs $x = (\nu, \epsilon)$.

stantly confirms several intuitive features about the class of Bayesian analyses, as well as providing clear quantitative statements in response to various robustness questions. We see that conditioning on ϵ and increasing ν always leads to a decrease in the posterior mean $\mathrm{E}[\theta|z,x]$, while conditioning on ν and increasing ϵ (and hence moving away from conjugacy) also decreases the mean. The experts may find it useful to know that moving away from conjugacy in this manner would lead to their posterior mean decreasing at most from 3.4 to approximately 2.75, as can be seen by drawing a vertical line above the blue dot, and that this mean is relatively insensitive to smaller likelihood contaminations of this form. A comparable lowering of the mean could also arise from choosing $\nu = 1.55$ instead of the original value of $\nu = 1$, showing that the analysis is far more sensitive to the prior standard deviation than to the likelihood contamination. For a careful interpretation, we should also take account of the emulator variance $\mathrm{Var}[f_1(x)|f_1(x_D)]$ and the corresponding credible intervals for $f_1(x)$ across $\mathcal{X}$, as is discussed for several example specifications in Section 2.2.

Figure 3 (right panel) shows the emulator expectation $\mathrm{E}[f_2(x)|f_2(x_D)]$ for the posterior standard deviation $f_2(x) = \mathrm{SD}[\theta|z,x]$ as a function of the inputs $x = (\nu, \epsilon)$, with the blue dot representing the original conjugate analysis. In contrast to the mean plot, this plot displays far more counterintuitive behaviour. Conditioning on ν and increasing ϵ has little effect for low ν and causes the SD to decrease monotonically for high ν. When $\epsilon = 0$ or 1, increasing ν leads to an increase in the posterior SD as expected. However, for intermediate values of the contamination ϵ, there are regions of $\mathcal{X}$ for which the opposite is true: an increase in the prior SD ν leads to a *decrease* in the posterior SD.

For example, a prior specification of $(\nu = 0.8, \epsilon = 0.72)$ has posterior SD = 0.50, but an increase in prior SD only to $(\nu = 2, \epsilon = 0.72)$ leads to a posterior SD = 0.46. So there are regions where being more certain a priori leads to one being comparatively less certain *a posteriori*. Note that this is not due to an over interpretation of the SD which may be too simple a summary of complex distributions, as exactly the same effect is seen, for example, when examining the width of the corresponding HPD intervals or the interquartile range. Nor is it an artefact of the emulation process, as has been checked by making further evaluations of the MCMC algorithm. Instead, this counter-intuitive behaviour can be explained in terms of a wider, less restrictive prior allowing the Bayesian update to be influenced by a larger range of the contaminated likelihood, sections of which may favour posteriors with lower variance.

Finding this non-trivial behaviour in a simple 1-dimensional case suggests that high-dimensional Bayesian analyses could easily exhibit similarly complex behaviour as we move away from conjugacy. The emulation methodology presented here is precisely designed to deal with high-dimensional cases of this form. Whether such complex behaviour was present would be difficult to discover without a careful global robustness/sensitivity analysis such as we propose here, which would be vital if the problem was deemed to be of high enough importance.

Example specifications

To further demonstrate the depth of analysis that is possible using Gaussian process emulation, we give the results of a small number of example specifications that could be provided by either single experts, or combinations of experts. We show that our analysis can give immediate and accurate answers in these cases, along with appropriate uncertainty statements that can be subsequently used to decide if further runs of the MCMC algorithm are required, to achieve a desired level of accuracy. We imagine that the following four specifications have been made:

Case 1 An expert has precise prior beliefs x_e corresponding to $\nu = 1.5$ and $\epsilon = 0.5$, but requests a local sensitivity analysis at this point.

Case 2 The experts have a fixed prior variance but want to explore the full range of contamination: $\nu = 0.8$, $0 \leq \epsilon \leq 1$.

Case 3 The experts have a fixed level of contamination, but imprecise prior variance such that: $0.5 \leq \nu \leq 1.9$, $\epsilon = 0.72$.

Case 4 The experts wish to perform a robustness analysis over a half elliptical region around the original conjugate analysis $(\nu = 1, \epsilon = 0)$ that satisfies

$$\frac{(\nu-1)^2}{0.3^2} + \frac{\epsilon^2}{0.4^2} \;<\; 1 \qquad \text{and} \qquad \epsilon > 0. \tag{23}$$

These four cases are shown in Figure 3 as the black cross and the black vertical, horizontal and curved lines respectively. The emulators derived in Section 2.2 can instantly provide the desired results for the four cases, as we now describe.

Table 2 gives the emulator expectation (first row) for the posterior mean $f_1(x) = \mathrm{E}[\theta|z,\nu,\epsilon]$ and SD, $f_2(x) = \mathrm{SD}[\theta|z,\nu,\epsilon]$ (first and forth columns) evaluated at the point $x_e = (1.5, 0.5)$, corresponding to the specification of case 1. As a local sensitivity analysis was requested, also given are the partial derivatives of $f_1(x)$ and $f_2(x)$ with respect to ν and ϵ, at this point, calculated as described in Section 2.1. These show that $f_1(x)$ is sensitive to both ν and ϵ at x_e; however, $f_2(x)$ is relatively insensitive to changes in ν. Most importantly, the second row of Table 2 gives the uncertainties due to the emulation process corresponding to each of these quantities, in the form of the emulator standard deviations, found from equation (8). These can be used to determine if a desired level of accuracy has been achieved, or if further MCMC runs are required.

	$f_1(x)$	$\frac{\partial f_1(x)}{\partial \nu}$	$\frac{\partial f_1(x)}{\partial \epsilon}$	$f_2(x)$	$\frac{\partial f_2(x)}{\partial \nu}$	$\frac{\partial f_2(x)}{\partial \epsilon}$
E[.]	2.596	−0.693	−0.423	0.533	−0.052	−0.243
SD[.]	0.020	0.113	0.179	0.006	0.028	0.049

Table 2: Results of the local sensitivity analysis corresponding to specification case 1. The first row gives the emulator expectation of all requested quantities of interest, namely the posterior mean $f_1(x)$, posterior SD $f_2(x)$ and partial derivatives of each. The second row gives the corresponding emulator SD of each of these estimates, which could be reduced using further MCMC runs.

Figure 4 shows the results for the posterior mean $f_1(x)$ (top left panel) and posterior SD $f_2(x)$ (bottom left panel) for the specification given in case 2, where here the contamination parameter ϵ varies along the x-axis. The blue lines give the emulator expectations, and the red lines give a 95% credible interval that represents the uncertainty due to the emulation process (and to a much smaller extent, due to the finite sample size of the MCMC draws). While both the posterior mean and SD appear to be monotonically decreasing with increasing ϵ, the posterior SD sharply decreases for $\epsilon > 0.7$. This alerts the expert to the fact that careful thought may be required when specifying levels of contamination above 0.7. Figure 4 gives the corresponding plots (top right: posterior mean, bottom right: posterior SD) for the specification given in case 3, where the prior standard deviation parameter ν varies along the x-axis. We can see that the posterior mean has been emulated to a high degree of accuracy compared to its variation over this range. The posterior SD exhibits some of the counterintuitive behaviour discussed in Section 2.2: once ν increases beyond approximately 0.8, the posterior SD decreases as a function of increasing ν. Here, the expert should be aware of both the counterintuitive behaviour and the sensitivity of the posterior to low values of ν. All four panels of Figure 4 also show multiple left-out diagnostic MCMC runs as black points. These were created by running the MCMC algorithm along the 1D regions defined by Case 2 and Case 3, and show good emulator performance. See Sup. Mat. for further diagnostics.

In many situations, the experts may purely want a robust Bayesian analysis performed over their specified regions $\mathcal{X}_k$ say, that is the identification of the maximum and minimum of the posterior quantities of interest over $\mathcal{X}_k$. For the maximum, we would hence wish to evaluate $\mathrm{E}[\max_{x_e \in \mathcal{X}_k} f(x_e)]$, where the expectation is performed over the Gaussian process, however, unlike all examples up to this point, we do not have an analytic expression for this term, as the distribution of the maxima and minima of

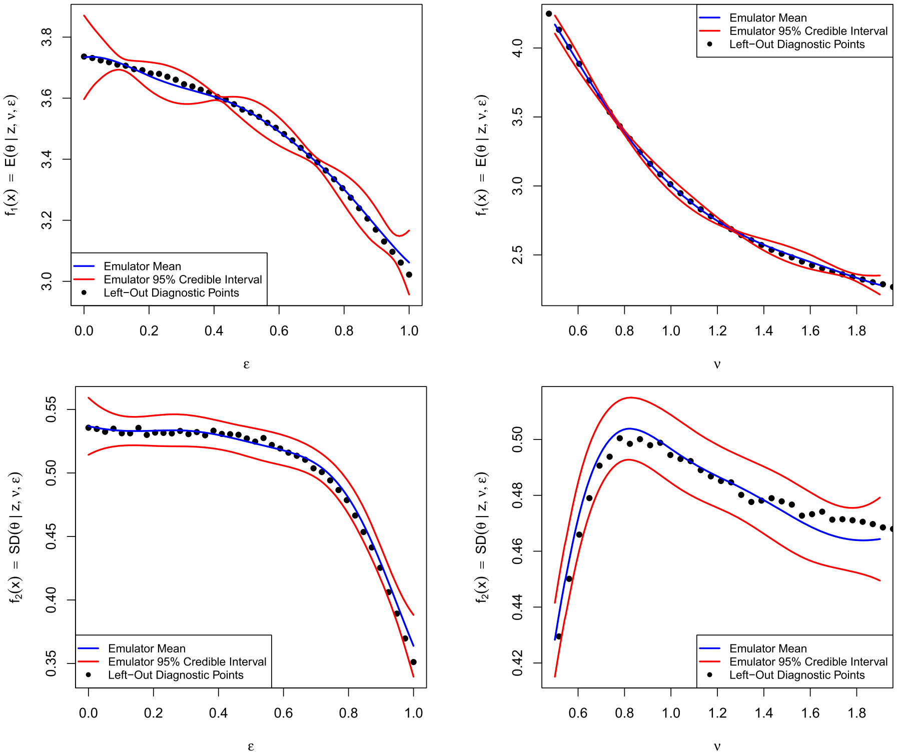

Figure 4: Emulator expectations (blue lines) and 95% credible intervals (red lines) for the expert specifications case 2 ($\nu = 0.8$, left column) and case 3 ($\epsilon = 0.72$, right column), with the results for the posterior means $\mathrm{E}[\theta|z,\nu,\epsilon]$ given by the top row, and the posterior standard deviations $\mathrm{SD}[\theta|z,\nu,\epsilon]$ by the bottom row. Black points are left-out diagnostic MCMC runs, showing good emulator performance.

a Gaussian process is only known for a small number of specific correlation functions. However we can easily approximate these expressions using simulation as follows, being careful to respect the smoothness of $f(x)$ and hence the joint structure of the emulator over $\mathcal{X}_k$. We define a large number of points $x_E^{(i)}$, $i = 1, \ldots, n_E$ spanning the specified region $\mathcal{X}_k$, and simulate jointly from the emulator across $x_E^{(i)}$. Specifically, we use the joint posterior distribution over the vector $f(x_E)$ of length n_E, which is given by

$$f(x_E)|f^{(s)}(x_D), m(.), c(.,.) \quad \sim \quad N(m^*(x_E), \Sigma^*(x_E)), \tag{24}$$

a direct consequence of equation (6), where $\Sigma^*(x_E)$ is a covariance matrix of dimension n_E, with elements $\Sigma^*_{ij} = c^*(x_E^{(i)}, x_E^{(j)})$. Equation (24) can be used to efficiently simulate a large number n_S of joint realisations from the posterior of the emulator. This provides,

$$f^{(j)}(x_E^{(i)}), \quad \text{with} \quad j = 1, \ldots, n_S \quad \text{and} \quad i = 1, \ldots, n_E \tag{25}$$

From these we may extract n_S maxima M_j and minima m_j and their corresponding means $\overline{M}$ and $\overline{m}$ respectively:

$$M_j = \max_i f^{(j)}(x_E^{(i)}) \,, \qquad\qquad m_j = \min_i f^{(j)}(x_E^{(i)}) \tag{26}$$

$$\overline{M} = \frac{1}{n_S}\sum_{j=1}^{n_S} M_j \,, \qquad\qquad \overline{m} = \frac{1}{n_S}\sum_{j=1}^{n_S} m_j \tag{27}$$

where $\overline{M}$ and $\overline{m}$ are estimates of $\mathrm{E}[\max_{x_e \in \mathcal{X}_k} f(x_e)]$ and $\mathrm{E}[\min_{x_e \in \mathcal{X}_k} f(x_e)]$ respectively.

Figure 5 (left panel) shows the estimated expected maxima $\overline{M}$ and minima $\overline{m}$ and the intervening range of the posterior mean $f_1(x)$ as the blue error bars, where $\overline{M}$ and $\overline{m}$ are the top and bottom of the blue error bars respectively, for each of the four cases, as labelled on the x-axis. The uncertainty due to the emulation process regarding these maxima and minima is represented by the red boxplots, which are formed from $n_S = 1000$ values of M_j and m_j respectively. Note the resulting asymmetries in some of the boxplots: e.g. the maxima of case 2: this is due to the correlation structure of the underlying GP calculation, which still respects the smoothness of the Bayesian analysis as a function of x. This can lead to accurate maximum and minimum estimates, even if the emulator uncertainty is high at individual input points. The blue points show the expected posterior means evaluated at the midpoint of the specification region, for each case, which give an approximate idea of any non-linearity of the posterior mean's response. The right-hand panel of Figure 5 shows the equivalent plot for the posterior standard deviations, $f_2(x)$. Once again, the emulator uncertainty, as represented by the red box-plots, shows how much variation could be resolved by further MCMC runs. In both panels, for cases 2 to 4, we can see that we have captured the majority of the variation of the robust Bayesian analysis (as given by the blue error bars), and that the emulator uncertainty is small in comparison, so it is unlikely that we would wish to design more MCMC runs. However, as a result of the correlation structure of the updated emulator, given as $c^*(x, x')$ in equation (8), the uncertain maximum and minimum of the GP may be correlated, and even possess a complex joint structure (see supplementary material for further discussion).

We now imagine that there is an important decision criteria that demands an alternative action if the posterior mean $f_1(x) < 2.6$ and the posterior standard deviation $f_2(x) < 0.47$ say. Experts in cases 1, 2 and 4 can rule out the alternative action immediately, as our analysis has confirmed that despite the imprecision in their specifications, their posteriors will not be close to the critical region. In case 3, these criteria are indeed possible, and the expert now knows that they need to think carefully about their original specification, particularly for the higher values of v where the critical region lies, as is confirmed by Figure 4 (right column). Should we need to do further exploration of the Bayesian analysis, in order to reduce the uncertainty about the location of such a critical region, we would perform additional waves of MCMC runs, using the well developed history matching methodology (see Vernon et al., 2010a,b, 2014).

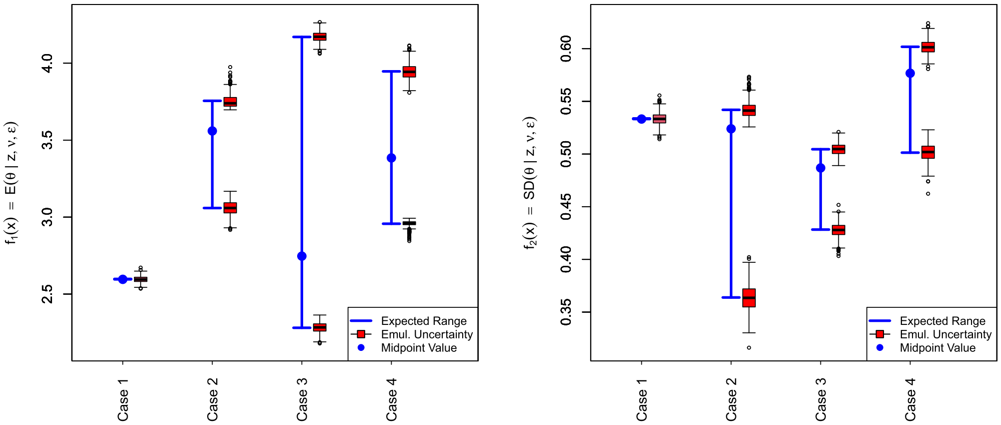

Figure 5: Output from the proposed analysis for the four example specifications given in Section 2.2. Left panel: the expected maximum and minimum of the posterior means $\mathrm{E}[\theta|z,\nu,\epsilon]$ are given by the top and bottom of the blue error bars. The uncertainty on these maximum and minimum estimates, due to the emulation process, is represented by the red box plots (based on 1000 realisations of the emulator), and could be reduced with further evaluations of the MCMC algorithm. The emulator expectation of the posterior mean at the midpoint of the imprecise specifications is given by the blue points. Right panel: the equivalent plot showing the possible ranges for the posterior standard deviations $\mathrm{SD}[\theta|z,\nu,\epsilon]$.

3 Application to a Bayesian analysis of river flow

3.1 Extension of a conjugate analysis

Vicens et al. (1975) give an account of a conjugate Bayesian analysis of annual streamflows of the Pemigewasset River at a measuring point at Plymouth, New Hampshire, USA. The data were the recorded flows in ft^3/s over the 60-year period of 1904-1963 (Survey, 2015). In their calculations, they assumed that the annual streamflows were identically and independently distributed as

$$z_i|\mu,\sigma^2 \quad \sim \quad \mathrm{N}\left(\mu,\sigma^2\right), \qquad i=1,\dots,60,$$

where μ and σ^2 were parameters that they wished to learn about. In order to have a conjugate analysis, the following prior specification for μ and σ was made:

$$\begin{aligned}\mu|\sigma^2 \quad &\sim \quad \mathrm{N}\left\{\mu_0, \left(\frac{\sigma}{n_0}\right)^2\right\},\\ \sigma^2 \quad &\sim \quad \text{Inv-Ga}(\alpha,\beta),\end{aligned}$$

where μ_0, n_0, α and β are hyperparameters that were specified.

We embed their analysis within a more general structure as follows. Because the data can be naturally thought of as a time series, the following simple extension can be made to the assumed data generating process:

$$\begin{aligned} z_i - \phi(z_{i-1} - \mu)|\mu, \sigma^2, \phi &\sim \mathrm{N}\left(\mu, \sigma^2\right), \qquad i = 2, \ldots, 60, \\ z_1|\mu, \sigma^2 &\sim \mathrm{N}\left(\mu, \sigma^2\right), \end{aligned}$$

where ϕ is a correlation parameter that could be fixed or we may be uncertain about. Because we are aiming to demonstrate just some of the utility of our approach and we have limited knowledge of the problem in hand, we will also investigate the following extension of the prior specification of Vicens *et al.* (1975):

$$\begin{aligned} \mu|\sigma^2 &\sim (1-\epsilon)\mathrm{N}_Q\left(Q_1, Q_3\right) + \epsilon\ \mathrm{C}_Q\left(Q_1, Q_3\right), \\ \sigma^2 &\sim \text{Inv-Ga}(\alpha, \beta), \end{aligned}$$

where Q_1 and Q_3 denote the lower and upper quartiles respectively and N_Q and C_Q are normal and Cauchy distributions that are parameterised using the lower and upper quartiles derived from

$$\mathrm{N}\left\{\mu_0, \left(\frac{\sigma}{n_0}\right)^2\right\}.$$

In order to complete this extended specification, we need to assign values to μ_0, n_0, α, β, ϕ and ϵ. Note that it would of course be possible to increase the sophistication of this structure, say by adding additional levels onto the underlying hierarchical model (e.g. by placing priors on ϕ, or indeed any of the parameters). However, as we wish to demonstrate our methods while also maintaining a clear comparison with the results of the original specification (and hence investigating its robustness), we choose to employ the robust analysis at this level. This has the benefit whereby the original specification can be identified as a single point within the space $\mathcal{X}$, which will aid interpretation.

When we use this prior specification with $\epsilon \neq 0$, we lose conjugacy and we need some numerical technique to derive the posterior distribution. For this application, we use a Metropolis-Hastings algorithm with proposal distributions:

$$\begin{aligned} \mu^*|\mu_{t-1}, \xi_\mu^2 &\sim N(\mu_{t-1}, \xi_\mu^2), \\ \sigma^{2*}|\sigma_{t-1}^2, \xi_\sigma^2 &\sim N(\sigma_{t-1}^2, \xi_\sigma^2). \end{aligned}$$

We use an adaptive algorithm to choose ξ_μ^2 and ξ_σ^2, and we use diagnostics to ensure the convergence of the Markov chains for each set of hyperparameters as in Section 2.2.

3.2 Emulation of the Bayesian analysis

We take as inputs to the computer model the specified parameters $x = (\mu_0, n_0, \alpha, \beta, \phi, \epsilon)$. We take as outputs $f(x)$ the posterior mean and variances of both μ and σ^2. The inputs and outputs are listed in Table 3 along with the ranges we decided to explore the analysis over that define the region $\mathcal{X}$. Note that the range for ϕ was chosen to be smaller than

Inputs x	Type of Input	Range	Outputs $f(x)$
μ_0	Prior hyperparameter	[500, 2000]	$\mathrm{E}[\mu\|z]$
n_0	Prior hyperparameter	[0.5, 30]	$\mathrm{Var}[\mu\|z]$
α	Prior hyperparameter	[100, 500]	$\mathrm{E}[\sigma^2\|z]$
β	Prior hyperparameter	[0, 30]	$\mathrm{Var}[\sigma^2\|z]$
ϕ	Autocorrelation parameter	[−0.2, 0.5]	
ϵ	Prior contamination	[0,1]	

Table 3: The inputs x and outputs $f(x)$ of the extended Bayesian analysis of river flow when represented as a computer model. The ranges for the inputs are also given to define the extent of the sensitivity analysis over $\mathcal{X}$.

the full $[-1, 1]$ range possible for correlations, as it was thought unlikely that an expert would assert such extreme values, due to meteorological considerations.

We create a 100-point design by creating a 99 point maximin Latin hypercube over the six dimensional hypercube $\mathcal{X}$ given by the ranges in Table 3 and adding a single input corresponding to the particular conjugate analysis carried out in Vicens et al. (1975). Again we choose to use slightly more runs than the rough guide suggestion of $10d$, to ensure reasonably accurate emulators after a single batch of runs. The parameters for the conjugate analysis were: $\mu_0 = 1,333$, $n_0 = 1$, $\alpha = 6.5$, $\beta = 402,057.5$, $\phi = 0$ and $\epsilon = 0$. We created a training set for our emulator by running the MCMC algorithm for each of the parameters and recording the four posterior moments of interest. The emulator was built using a Matérn correlation function, a linear mean function and an extra variance term to capture variability in the MCMC estimation process as described in Section 2.2. We also checked the performance of the emulator using the diagnostic tools of Bastos and O'Hagan (2008), and found that the uncertainty caused by employing an emulator was generally two orders of magnitude smaller than the range of different values we observed for each of the four outputs of interest.

Figure 6 shows the effect of changing some of the parameters for three of the outputs of interest. The red line in each plot gives the average value for the output named on the y-axis conditional on the fixed value of the input from the x-axis. For these sensitivity analysis plots, we assume uniform distributions over all the ranges given in Table 3. The grey regions on the plot show a 90% credible interval for the different outputs conditional on the fixed input value and can be thought of being illustrative of plausible values for the output given the fixed input value. The top-right plot of Figure 6 shows that as we vary α the posterior mean of μ will on average stay at 1347 ft^3/s, but the plausible range of values shrinks slightly as we increase α. We can of course create such plots for each input-output combination, and the four shown in Figure 6 are the most interesting for this example in that the ranges and mean change over the range of the input. The top-left plot of Figure 6 shows the potentially unexpected effect that changing μ_0 has on the posterior mean of μ in this analysis: relatively small deviations from the original specification of μ_0 can have a large effect on the posterior mean of μ. This information is obviously useful to any interested in the robustness and sensitivity of this hyperparameter in this analysis. The fact that the posterior mean is, on average,

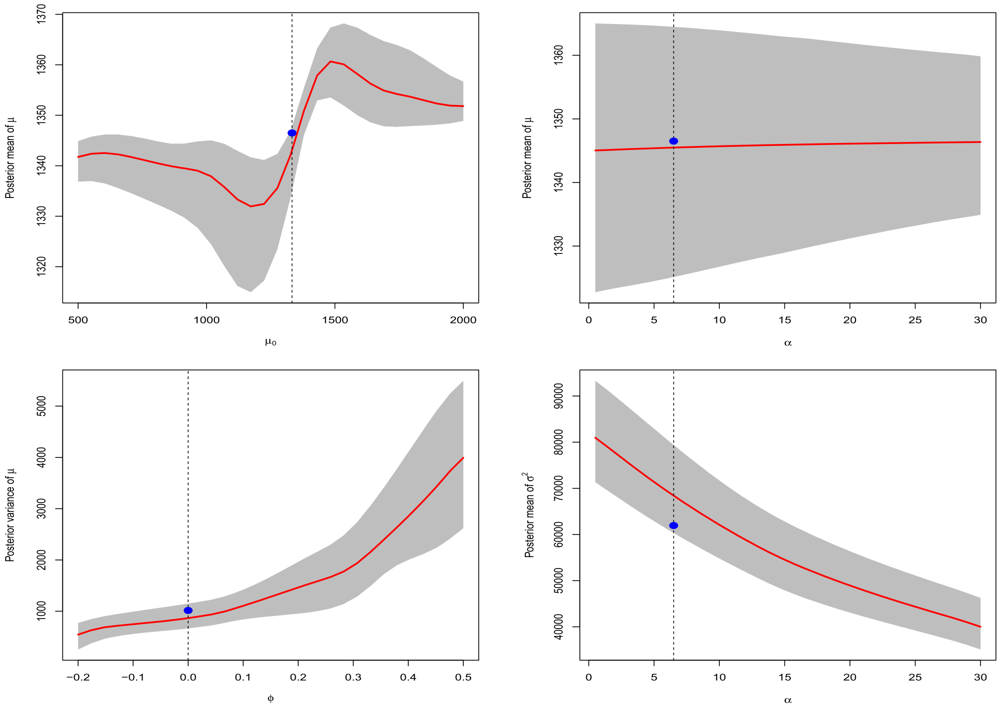

Figure 6: Main effects plots, showing the average effect of inputs μ_0, α and ϕ on various outputs (red line). The grey envelope represents the possible output values (due to varying the other five inputs) as a 90% credible interval, conditioned on the given input. The blue points give the outputs corresponding to the prior specification for the parameters made by Vicens et al. (1975).

stable for values of μ_0 below 1,000 and above 1,700 would be of interest to scientists who have prior beliefs that accord with one of those possibilities in that they will know that relatively little effort should be spent on eliciting their beliefs about μ_0 precisely. This behaviour is due to the Cauchy part of the prior distribution dominating the Bayesian update, when there is modest prior-data conflict. We must of course interpret such plots with caution, and may choose to further investigate interesting regions (for example, where $\epsilon \to 0$) with a second wave of MCMC runs, as discussed in Section 4.2.

The variance contributions of each of the inputs to each of the outputs of interest are calculated to show the influence of each input using the probabilistic sensitivity analysis method of Oakley and O'Hagan (2004) (again, using uniform distributions over the ranges in Table 3). The results are given in Table 4. In the table, the main-effect index column shows the percentage of variance in the output that is due to the corresponding input alone, and the total-effect index is the percentage of variance that is due to the corresponding input and all of the higher-order interactions it is involved in (Saltelli et al., 2000). Immediately, from the table, we can see which parameters have

	$\mathrm{E}(\mu\|z)$		$\mathrm{Var}(\mu\|z)$		$\mathrm{E}(\sigma^2\|z)$		$\mathrm{Var}(\sigma^2\|z)$	
	Main-effect index(%)	Total-effect index(%)	Main-effect index(%)	Total-effect index(%)	Main-effect index(%)	Total-effect index(%)	Main-effect index(%)	Total-effect index(%)
μ_0	71	99	0	0	0	0	0	0
n_0	0	4	0	4	0	0	0	1
α	0	2	11	21	85	87	88	93
β	0	0	1	5	5	6	2	4
ϕ	1	24	75	85	8	9	5	8
ϵ	0	1	0	5	0	1	0	1

Table 4: Variance-based sensitivity indices for the μ outputs.

most impact on the different outputs of interest: for instance, we can see that (over the ranges specified) μ_0 is accounting for the majority of the variation in the output $\mathrm{E}(\mu|z)$ as expected, but μ_0 is having no discernible effect on any other output of the analysis. The impact of the contamination parameter ϵ across all the analysis outputs can also be seen to be relatively small, which may be of interest to any person who questioned the choice of the normal prior in the original paper. From Table 4, we can also see that the input ϕ is having an effect on $\mathrm{E}(\mu|z)$, but only in interaction with the other input parameters (most probably μ_0). Given this information, we may want to investigate the changes in $\mathrm{E}(\mu|z)$ when we jointly manipulate μ_0 and ϕ.

In addition to the plots in Figure 6, we are able to visualise the joint effect of two inputs by plotting the average value of the outputs conditioning on fixed values of two of the inputs. The joint effect of μ_0 and ϕ on $\mathrm{E}[\mu|z]$ is shown in the plot of Figure 7: it is clear from that plot that the level of autocorrelation ϕ, changes the influence of μ_0, with larger positive values of ϕ resulting in a much stronger dependence on μ_0. Again, these types of plots can be used to identify regions of the input space where the analysis is robust. Like for the plots of Figure 6, we could have presented these plots for any input-pair and output combination, but, for the most part, these plots were either flat, or just showed interesting behaviour in one dimension (which is represented by Figure 6).

On all the plots in Figures 6 and 7, the location of the result of the original analysis from Vicens et al. (1975) is shown as a blue dot. By considering the plots and the variance-based sensitivity analysis results we can judge which outputs are robust to changes in which inputs within the vicinity of the original analysis. For instance, it is clear from these results that careful consideration needs to be given to the specification of μ_0, ϕ and α, and we know that the output $\mathrm{E}(\mu|z)$ is particularly sensitive to changes in μ_0 around the value used in the original analysis.

The emulator can also be used in a predictive manner: another scientist may come along who agrees with the original specification of μ_0, n_0, α and β, but they believe that there is autocorrelation that is captured by setting $\phi = 0.25$ and that the prior should be Cauchy rather than normal ($\epsilon = 1$). The emulator can be queried to find immediately that, under this specification, we have the results in Table 5, where we have an estimate of the relevant Bayesian analysis and an appreciation of the uncertainty caused by

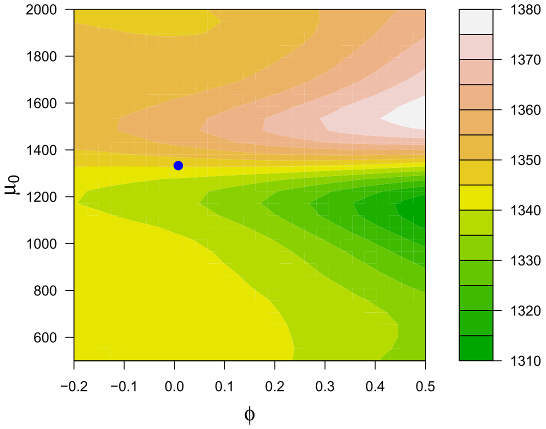

Figure 7: Joint effects plot showing the average effect of pairs of the inputs ϕ and μ_0 on the analysis output of $\mathrm{E}[\mu|z]$.

Posterior summary	Results from Vicens et al. (1975)	Median	90% credible interval
$\mathrm{E}[\mu\|z]$	1,347	1,344	(1,339,1,350)
$\mathrm{Var}[\mu\|z]$	1,015	1,749	(1,424,2,073)
$\mathrm{E}[\sigma^2\|z]$	61,937	68,210	(67,420,68,990)
$\mathrm{Var}[\sigma^2\|z]$	1.11×10^8	1.47×10^8	$(1.32\times10^8,1.62\times10^8)$

Table 5: Predictions from the emulator when using original values for μ_0, n_0, α and β alongside $\phi = 0.25$ and $\epsilon = 1$ (all results are to four s.f.).

approximating the analysis using the emulator. We can see in this particular case that the posterior mean for μ is in the vicinity of the value from the original analysis (1,347), but the posterior variance for μ is far greater than the original (1,015), which is to be expected due to the increased correlation in the likelihood giving a reduced effective sample size, combined with a more uncertain prior. If this scientist was uncertain about the level of autocorrelation and wished to specify a distribution for ϕ, the emulator could still be used to find their posterior summaries using the usual uncertainty analysis approach of Oakley and O'Hagan (2002).

Of course, considering the sensitivity of the outputs in this way and performing predictions based on the emulator are just two of the ways we can investigate the original analysis using our method: we can also perform the type of analyses that were covered in Section 2.2 and many more as discussed later in the present article. In particular, if there was interest in further exploring the robustness of the analysis in a particular part of input space, we could use the emulator to help select further MCMC runs for different inputs in order to increase our knowledge of how the posterior summaries are affected in that region. The emulator-building exercise and the subsequent plotting of main and joint effects can also be useful as a diagnostic tool in that we may have prior beliefs about the way in which inputs influence the posterior outputs and these plots can help identify unexpected behaviour that may be due to programming bugs in the MCMC implementation.

4 Discussion and future research directions

4.1 Modelling choices

The proposed methodology raises many questions, some of which are related to those seen in a standard robust Bayesian analysis. We give our thoughts on some of the key issues as follows (see Insua and Ruggeri (2000) for further discussion):

- **What parts of the prior and likelihood should we vary?** When setting up such analyses, we have many choices over what parts of the prior and likelihood we could vary. These can in the first instance be guided by the differing views held within the scientific community, or by our desire to test the sensitivity of our analysis to important parts of the specification. Key elicited quantities (such as the prior variance ν in Section 2.2) are obvious choices for parameters to vary, as are parameters representing relatively arbitrary but convenient assumptions, such as the additive contamination parameter ϵ from both Sections 2.2 and 3, which breaks conjugacy for $\epsilon > 0$, or indeed multiplicative contaminations of the form $p^{(1-\epsilon)}q^{\epsilon}$ motivated say from a differential geometry perspective. In full generality, we may wish to vary everything possible, while maintaining consistency with the limited prior and likelihood specification. However, we should be careful here as although not explicitly stated, the prior specification may contain further reasonable but implicit structural information, such as unimodality and continuity of both the prior and likelihood pdfs, as well as additional, and possibly quite strict, bounds on the derivatives of the pdf's to ensure smoothness (the expert's beliefs, were we to interrogate them further, are unlikely to be jagged). This is critical as many robust analyses that leave out such additional constraints, can produce relatively non-informative results, especially in high dimension (this links to arguments made by Gustafson and Wasserman, 1995). These constraints therefore greatly restrict the class of analyses we should use, and hence may allow parameterised approaches, such as we present here, to capture the major sources of variation. The limitations as to what we can vary link to the concept of *model discrepancy* that we discuss in Section 4.2, which would capture the additional uncertainty that our current representation ignores.

- **How do we decide how to contaminate a prior or likelihood?** As we have demonstrated, the contamination of a prior or likelihood represents a simple to implement parameterised method of breaking away from mathematically convenient distributional assumptions, while respecting core scientific principles. There is of course much freedom in the choice of contaminating distribution, however, we would usually want to ensure the contaminant possesses key attributes found in the uncontaminated term. For example, in Section 2.2 the contaminant to the likelihood $\pi(z_i|\theta, \epsilon = 1)$ given in equation (17) was chosen to have the same expectation of $1/\theta$, as the uncontaminated term $\pi(z_i|\theta, \epsilon = 0)$ given by equation (14), but note that this is not a restriction or requirement of our approach, just a specific choice we made to mimic say an expert adhering to that physical principle. Similarly the prior contamination of Section 3.1, shared the same quartiles as the original prior specification. Note that in both cases the contaminants shared with the original analysis the additional properties of unimodality and continuity of the pdfs (and derivatives). While these perturbations are by their very nature limited, it would still be comforting to find that the key features of the posterior, or end decision process, are robust to them, and highly informative to find the opposite.

- **What should the exploratory space $\mathcal{X}$ look like? In the simplest case, what ranges should we use?** Ideally, the exploratory space $\mathcal{X}$ should contain the differing specifications that exist across all, or at least some specific subset, of the relevant scientific community. Note that $\mathcal{X}$ may contain regions $\mathcal{X}_k$ representing individual robust Bayesian analyses that scientists wish to perform, such as cases 2, 3 and 4 in Section 2.2. While in practice this would be difficult to achieve precisely, as the scientific community may not agree exactly with our choice of parameterisation, we would still hope to capture the major aspects of the differences of opinion across the area. This implies that there is an important difference between $\mathcal{X}$ and the corresponding region $\mathcal{X}_k$ explored in a single perfunctory robust Bayesian analysis, in that $\mathcal{X}$ just needs to cover all areas of interest, and assuming it achieves this, the precise location of its boundary is of somewhat less importance (however, although in principal our proposed emulation methodology can deal with large numbers of inputs defined over wide ranges, as discussed below, the smaller $\mathcal{X}$ is, the easier it may be to emulate). In contrast, when specifying a particular region $\mathcal{X}_k$ for use in a robust Bayesian analysis, where interest lies in the extrema of $f(x)$ over $\mathcal{X}_k$, such as in cases 2 to 4 in Section 2.2, the geometry and extent of the boundaries of $\mathcal{X}_k$ should be considered very carefully. For example, often, $\mathcal{X}_k$ may be constructed from the intersection of univariate interval constraints on the components of x, implying $\mathcal{X}_k$ is a hypercube. However, this is usually just a convenient construction, and can possess disadvantages: as $f(x)$ may display noticeably different behaviour in the many corners of such a hypercube, the corners may dominate the robust analysis. An elliptical specification, as used in case 4 in Section 2.2, may be both more realistic and simultaneously easier to emulate. These issues have caused problems in previous robustness studies in differing dimensions: while exploring wide classes of priors in 1-dimension can still lead to meaningful conclusions (Berger, 1994), in higher dimensions such artificial classes of priors can overwhelm the data, leading to non-informative results (Insua and Ruggeri, 2000), a problem that will become worse when we simultaneously perturb the likelihood. We see that the requirement to specify $\mathcal{X}$ is not created by our analysis, rather it already exists for *any* robust Bayesian analysis and by extension for *any* Bayesian analysis. Our approach just helps one to explore $\mathcal{X}$ in a principled manner and should help facilitate the analysis of quite large spaces, providing deeper insight.

- **What happens if we cannot emulate the Bayesian analysis?** One can envisage a particularly erratically behaved Bayesian analysis where standard emulation procedures would perform poorly, as would most likely be flagged by emulator diagnostics (Bastos and O'Hagan, 2008). In this case, we would a) be very glad to have been made aware of this erratic behaviour across $\mathcal{X}$ and b) most likely suggest the analysis would fail any reasonable test of robustness. Hence if we cannot emulate it, we would be unlikely to trust it. We may then attempt to emulate sub-components of the full analysis, to investigate its structure further and to identify the cause of the non-robust behaviour. One possible cause of such erratic MCMC output could be that the original Bayesian analysis suffers from identifiability issues (Gustafson, 2015), again which may be picked up here by emulator diagnostics. Other challenges can arise if the original Bayesian analysis is both complex and we wish to perturb many of its attributes, for example in the case of complex Bayesian linear mixed models, or various spatio-temporal models. This may lead to the dimension of $\mathcal{X}$ being very large, possibly requiring substantial numbers of MCMC evaluations, which despite being embarrassingly parallelizable, may still be impractical to perform. While GP emulation can handle moderate to high-dimensional functions, this requires more sophisticated emulator forms, e.g. as shown in equation (7) in the Sup. Mat., which are designed to exploit both global smoothness, and dimension reduction strategies in the form of active/inactive inputs. In addition, if the dimension of $\mathcal{X}$ is still too large, we may want to carefully consider limiting our analysis to interesting subspaces of $\mathcal{X}$ that we suspect will drive most of the behaviour of $f(x)$, perhaps those aligned with the main sources of disagreement between domain experts' specifications. Here again, emulation may be successful or would at least highlight classes of features that require further investigation. See Sup. Mat. for further discussion.

- **What can we do if we cannot assume smoothness?** If we are uncomfortable with the standard smoothness assumption across $\mathcal{X}$, we can use alternative forms for the emulator correlation function that represent non-smooth surfaces with particular attributes. If we suspect the output to have sudden discontinuities, either in its derivative or in the function $f(x)$ itself, we can attempt to identify the location of such discontinuities using history matching techniques discussed below.

4.2 Future research directions

The proposed methodology raises several future research directions, some of which we highlight here (see the supplementary material for details). For example, there are powerful computer model techniques that have interesting analogies in this context:

- **History matching**: say interest lies in identifying a subset $\mathcal{X}_0 \subset \mathcal{X}$ that satisfies some criteria on the posterior, possibly related to a downstream decision calculation, or related to finding regions of high sensitivity. In this case we can employ the computer model technique of history matching: a global search strategy that efficiently exploits the structures of the emulators using iterative waves of runs (e.g. see Vernon et al., 2010a,b; Rodrigues et al., 2017; Williamson et al., 2013; Andrianakis et al., 2017).

- **Model discrepancy**: a key feature of current computer model analyses is the inclusion of a model discrepancy term (Craig et al., 1997; Kennedy and O'Hagan, 2001; Goldstein et al., 2013), an upfront acknowledgement of the deficiencies of a scientific computer model due to missing physics, simplifying assumptions, imperfect solvers etc. In our current context of a Bayesian analysis, the model discrepancy would represent the uncertainty due to the simplifying assumptions used throughout the construction of the Bayesian model and prior specification, beyond those explored by the robust analysis itself. It would therefore link our current robust analysis with the robust analysis that we would wish to do given more time, computational resources and expert input.

Other lines of research more specific to this context are also possible:

- **Structured Emulator Priors**: There are several situations where we would have detailed insight into the result of the Bayesian update for specific subsets of $\mathcal{X}$, for example, there may be surfaces in $\mathcal{X}$ where we can solve the Bayesian update exactly due to certain prior variances equalling zero, or due to conjugacy. Other information may be provided by fast, but approximate algorithms. Depending on the posterior features of interest, there is a rich hierarchy of informed priors one could use that incorporates such information for minimal computational cost. Here, we would elicit priors from the statistician, not the subject matter expert, hence detailed elicitations may be possible.

- **MCMC development**: While the use of MCMC in our approach is trivially parallelisable, we also envisage improvements to MCMC algorithms tailored specifically to this type of analysis, that exploit the geometry of $\mathcal{X}$. For example, the final end state and tuning parameters values for a chain located at $x_1 \in \mathcal{X}$ would make an ideal initial condition for a chain located at $x_2 \in \mathcal{X}$ were $|x_1 - x_2|$ considered small. Perhaps of most use would be the incorporation of MCMC based local sensitivity analysis (Perez et al., 2005) into the emulator via the derivative structure given by equation (12), which may lead to substantial improvement in emulator accuracy.

5 Conclusion

In this article we have proposed a framework for addressing the general Bayesian robustness problem, in which we treat complex and computationally demanding Bayesian analyses as expensive computer models. We applied emulation technology developed for complex computer models to explore the structure of the Bayesian analysis itself, and, specifically, its response to various changes in both the prior and likelihood specification. This allows for a more general sensitivity and robustness analysis, and provides a very flexible methodology that could also be applied to a wide class of statistical analyses.

It could be argued that every important Bayesian analysis, where the results may have serious consequences, should employ a robust analysis of the kind we propose here. Enabling the analysis of classes of prior and likelihood specification should also help the uptake of Bayesian methods within scientific communities, as each expert will have access to the posterior attributes corresponding to their own subjective beliefs. Experts may also find the answer to the question: “how far do I have to perturb my specification before my decision changes” to be easy to interpret, and help assuage their fears over the use of specific choices of subjective priors and likelihood, due to their (possible) robustness. In many contexts this strategy may be of more use that the standard Bayesian approach of adding another layer to the model hierarchy, requiring the assertion of possibly artificial priors over the space $\mathcal{X}$, the meaning of which may be questionable.

There are now some extremely expensive MCMC algorithms that may take weeks or longer to run, and hence there may be obvious practical challenges that make it impossible to perform repeated evaluations as required for this analysis. However, in many of these cases other, faster but approximate Bayesian methods will be available such as Hamiltonian Monte Carlo, Variational Inference or even Approximate Bayesian Computation. We can apply our emulation based robust analysis to these approximations either directly, or by incorporating them within a multilevel analysis (see Sup. Mat. for further discussion). In the case where the MCMC is indeed too slow for our analysis, and no reliable alternative approximations are available, we should insist on turning the argument around, and our response would be the same as is often given to climate scientists, who also construct extremely expensive models: if the model is too expensive to allow a reasonable sensitivity analysis, why should we trust in its results at all? Such considerations would promote a welcome change in emphasis in Bayesian statistics away from extremely complex models and algorithms, and toward well understood, robust and trustworthy analyses. This, combined with further developments to tailor MCMC and other approximate algorithms for efficient use in this context, maybe a sensible direction for Bayesian statistics to take.

Supplementary Material

A Bayesian computer model analysis of Robust Bayesian analyses: Supplementary Material (DOI: 10.1214/22-BA1340SUPP; .pdf). Further details regarding a) an introduction to computer model emulation, b) the example Bayesian model: emulator diagnostics, posterior quantile emulation, c) comparison with importance sampling, d) extended discussion of future research directions.

References

Aczel, B., Hoekstra, R., and Gelman, A. e. a. (2020). "Discussion points for Bayesian inference." *Nature Human Behaviour*, 4: 561–563. 3

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T., Oakley, J., Nsubuga, R., Goldstein, M., and White, R. (2015). "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda." *PLoS Computational Biology*, 11(1): e1003968. 5, 7

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2017). "History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(4): 717–740. MR3670414. doi: https://doi.org/10.1111/rssc.12198. 7, 26

Bastos, L. and O'Hagan, A. (2008). "Diagnostics for Gaussian process emulators." *Technometrics*, 51: 425–38. MR2756478. doi: https://doi.org/10.1198/TECH.2009.08019. 6, 20, 25

Berger, J. O. (1994). "An overview of robust Bayesian analysis." *Test*, 3(1): 5–59. MR1293110. doi: https://doi.org/10.1007/BF02562676. 1, 2, 25

Berger, J. O., Insua, D. R., and Ruggeri, F. (2000). "Bayesian Robustness." In Insua, D. R. and Ruggeri, F. (eds.), *Robust Bayesian Analysis*, Lecture Notes in

Statistics, chapter 1, 1–31. Springer. MR1795207. doi: https://doi.org/10.1007/978-1-4612-1306-2_1. 1

Berger, J. O. and Sellke, T. (1987). "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." *Journal of the American Statistical Association*, 82(397): 112–122. URL http://www.jstor.org/stable/2289131 MR0883340. 2

Box, G. E. and Tiao, G. C. (1962). "A further look at robustness via Bayes's theorem." *Biometrika*, 49(3-4): 419–432. MR0157447. doi: https://doi.org/10.1093/biomet/49.3-4.419. 1

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press. MR2742422. doi: https://doi.org/10.1201/b10905. 11

Chipman, H., Ranjan, P., and Wang, W. (2012). "Sequential design for computer experiments with a flexible Bayesian additive model." *The Canadian Journal of Statistics*, 40(4): 663–678. MR2998855. doi: https://doi.org/10.1002/cjs.11156. 6

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997). "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion)." In Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D. (eds.), *Case Studies in Bayesian Statistics*, volume 3, 36–93. New York: SV. MR1425400. 4, 5, 6, 26

Dey, D. and Micheas, A. (2000). "Ranges of Posterior Expected Losses and ϵ-Robust Actions." In Insua, D. R. and Ruggeri, F. (eds.), *Robust Bayesian Analysis*, Lecture Notes in Statistics, chapter 8, 145–160. Springer. MR1795214. doi: https://doi.org/10.1007/978-1-4612-1306-2_8. 2

Edwards, T. L., Nowicki, S., and Marzeion, B. e. a. (2021). "Projected land ice contributions to twenty-first-century sea level rise." *Nature*, 593: 74–82. 5

Fan, T. H. and Berger, J. (2000). "Robust Bayesian displays for standard inferences concerning a normal mean." *Computational Statistics and Data Analysis*, 33(4): 381–399. 2, 5

Francom, D., Sansó, B., Kupresanin, A., and Johannesson, G. (2018). "Sensitivity analysis and emulation for functional data using bayesian adaptive splines." *Statistica Sinica*, 28(2): 791–816. MR3791088. 6

Geweke, J. (1999). "Simulation Methods for Model Criticism and Robustness Analysis." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, volume 6. Oxford University Press. MR1723501. 5

Geyer, C. J. (1994). "Estimating Normalizing Constants and Reweighting Mixtures." Technical report, University of Minnesota. 5

Giordano, R., Broderick, T., and Jordan, M. I. (2018). "Covariances, Robustness, and Variational Bayes." *Journal of Machine Learning Research*, 19: 1–49. MR3874159. 2

Goldstein, M., Seheult, A., and Vernon, I. (2013). *Environmental Modelling: Finding*

Simplicity in Complexity, chapter Assessing Model Adequacy. Chichester, UK: John Wiley & Sons, Ltd, second edition. 26

Grzeszczuk, R., Terzopoulos, D., and Hinton, G. (1998). "Neuroanimator: Fast neural network emulation and control of physics-based models." In *SIGGRAPH'98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 9–20. Association for Computing Machinery. 6

Gu, M. and Berger, J. O. (2016). "Parallel Partial Gaussian Process Emulation for Computer Models with Massive Output." *Annals of Applied Statistics*, 10(3): 1317–1347. MR3553226. doi: https://doi.org/10.1214/16-AOAS934. 5

Gustafson, P. (2000). "Local robustness in Bayesian analysis." In Insua, D. R. and Ruggeri, F. (eds.), *Robust Bayesian Analysis*, Lecture Notes in Statistics, chapter 4, 71–88. Springer. MR1795210. doi: https://doi.org/10.1007/978-1-4612-1306-2_4. 2

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models*. Chapman and Hall/CRC, 1st edition. 26

Gustafson, P. and Wasserman, L. (1995). "Local Sensitivity Diagnostics For Bayesian Inference." *The Annals of Statistics*, 23(6): 2153–2167. MR1389870. doi: https://doi.org/10.1214/aos/1034713652. 2, 24

Heitmann, K., Higdon, D., et al. (2009). "The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum." *Astrophys. J.*, 705(1): 156–174. 5

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004). "Combining field data and computer simulations for calibration and prediction." *SIAM Journal on Scientific Computing*, 26(2): 448–466. MR2116355. doi: https://doi.org/10.1137/S1064827503426693. 4

Insua, D. R. and Ruggeri, F. (eds.) (2000). *Robust Bayesian Analysis*. Lecture Notes in Statistics. Springer. MR1795206. doi: https://doi.org/10.1007/978-1-4612-1306-2. 2, 24, 25

Johnson, J. S., Gosling, J. P., and Kennedy, M. C. (2011). "Gaussian process emulation for second-order Monte Carlo simulations." *Journal of Statistical Planning and Inference*, 141: 1838–48. MR2763214. doi: https://doi.org/10.1016/j.jspi.2010.11.034. 7

Kallioinen, N., Paananen, T., Bürkner, P.-C., and Vehtari, A. (2021). "Detecting and diagnosing prior and likelihood sensitivity with power-scaling." ArXiv:2107.14054 [stat.ME]. 2

Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian calibration of computer models (with Discussion)." *Journal of the Royal Statistical Society. Series* B, 63: 425–64. MR1858398. doi: https://doi.org/10.1111/1467-9868.00294. 4, 5, 6, 26

Liu, F. and West, M. (2009). "A dynamic modelling strategy for Bayesian computer model emulation." *Bayesian Analysis*, 4(2): 393–411. URL https://doi.org/10.1214/09-BA415 MR2507369. doi: https://doi.org/10.1214/09-BA415. 6

Loeppky, J., Sacks, J., and Welch, W. (2009). "Choosing the Sample Size of a Computer Experiment: A Practical Guide." *Technometrics*, 51: 366–376. MR2756473. doi: https://doi.org/10.1198/TECH.2009.08040. 6

Moreno, E. (2000). "Global Bayesian Robustness for Some Classes of Prior Distributions." In Insua, D. R. and Ruggeri, F. (eds.), *Robust Bayesian Analysis*, Lecture Notes in Statistics, chapter 3, 45–70. Springer. MR1795209. doi: https://doi.org/10.1007/978-1-4612-1306-2_3. 2

Morris, M. D. and Mitchell, T. J. (1995). "Exploratory designs for computational experiments." *Journal of statistical planning and inference*, 43(3): 381–402. 7

Muller, U. (2012). "Measuring prior sensitivity and prior informativeness in large Bayesian models." *Journal of Monetary Economics*, 59(6): 581–597. 2

Oakley, J. and O'Hagan, A. (2002). "Bayesian inference for the uncertainty distribution of computer model outputs." *Biometrika*, 89(4): 769–784. MR2088780. doi: https://doi.org/10.1111/j.1467-9868.2004.05304.x. 23

Oakley, J. E. (2009). "Decision-theoretic sensitivity analysis for complex computer models." *Technometrics*, 51(2): 121–129. MR2668169. doi: https://doi.org/10.1198/TECH.2009.0014. 4

Oakley, J. E. and O'Hagan, A. (2004). "Probabilistic sensitivity analysis of complex models: a Bayesian approach." *Journal of the Royal Statistical Society. Series* B, 66: 751–69. MR2088780. doi: https://doi.org/10.1111/j.1467-9868.2004.05304.x. 6

O'Hagan, A. (1992). "Some Bayesian numerical analysis." In *Bayesian Statistics 4,* Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 345–63. Oxford: Oxford University Press. MR1380285. 6, 8

O'Hagan, A. (2006). "Bayesian analysis of computer code outputs: a tutorial." *Reliability Engineering & System Safety*, 91: 1290–300. 2

Perez, C., Martin, J., and Rufo, M. J. (2005). "MCMC-based local parametric sensitivity estimations." *Computational Statistics and Data Analysis*, 51(2): 823–835. MR2297491. doi: https://doi.org/10.1016/j.csda.2005.09.005. 2, 27

Peruggia, M., Santner, T. J., and Ho, Y. Y. (2004). "Detecting stage-wise outliers in hierarchical Bayesian linear models of repeated measures data." *Annals of the Institute of Statistical Mathematics*, 56(3): 415–433. MR2095011. doi: https://doi.org/10.1007/BF02530534. 3

Robert, C. P. and Rousseau, J. (2016). "Nonparametric Bayesian Clay for Robust Decision Bricks." *Statistical Science*, 31(4): 506–510. URL http://dx.doi.org/10.1214/16-STS567 MR3598730. doi: https://doi.org/10.1214/16-STS567. 2

Rodrigues, L. F. S., Vernon, I., and Bower, R. G. (2017). "Constraints to galaxy formation models using the galaxy stellar mass function." *MNRAS*, 466(2): 2418–2435. 26

Roos, M. and Held, L. (2011). "Sensitivity analysis in Bayesian generalized linear mixed models for binary data." *Bayesian Analysis*, 6(2): 259–278. MR2806244. doi: https://doi.org/10.1214/11-BA609. 2

Roos, M., Martins, T., Held, L., and Rue, H. (2015). "Sensitivity analysis for Bayesian hierarchical models." *Bayesian Analysis*, 10(2): 321–349. MR3420885. doi: https://doi.org/10.1214/14-BA909. 1

Saltelli, A., Chan, K., and Scott, E. (eds.) (2000). *Sensitivity Analysis*. New York: Wiley. MR1886391. 6, 21

Santner, T., Williams, B., and Notz, W. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer. MR2160708. doi: https://doi.org/10.1007/978-1-4757-3799-8. 6

Shyamalkumar, N. D. (2000). "Likelihood Robustness." In Insua, D. R. and Ruggeri, F. (eds.), *Robust Bayesian Analysis*, Lecture Notes in Statistics, chapter 7, 127–143. Springer. MR1795213. doi: https://doi.org/10.1007/978-1-4612-1306-2_7. 2

Sinharay, S. and Stern, H. S. (2002). "On the Sensitivity of Bayes Factors to the Prior Distributions." *The American Statistician*, 56(3): 196–201. MR1940207. doi: https://doi.org/10.1198/000313002137. 5

Smith, A. F. M. and Gelfand, A. E. (1992). "Bayesian Statistics without tears: A Sampling-Resampling Perspective." *The American Statistician*, 46(2): 84–88. MR1165566. doi: https://doi.org/10.2307/2684170. 5

Springer Nature (2022). "Search results for query "Bayesian" for nature journal articles." URL https://www.nature.com/search?q=bayesian&journal=nature&article_type=research&order=relevance&title=Bayesian 1

Survey, U. G. (2015). "USGS 01076500 Pemigewasset River at Plymouth, NH." URL http://waterdata.usgs.gov/nwis/inventory/?site_no=01076500&agency_cd=USGS 18

Vernon, I. and Goldstein, M. (2022). "Bayes linear emulation and history matching of stochastic systems biology models." *In Preparation*. 7

Vernon, I., Goldstein, M., and Bower, R. G. (2010a). "Galaxy Formation: a Bayesian Uncertainty Analysis." *Bayesian Analysis*, 5(4): 619–670. MR2740148. doi: https://doi.org/10.1214/10-BA524. 4, 5, 6, 12, 17, 26

Vernon, I., Goldstein, M., and Bower, R. G. (2010b). "Rejoinder for Galaxy Formation: a Bayesian Uncertainty Analysis." *Bayesian Analysis*, 5(4): 697–708. MR2221263. doi: https://doi.org/10.1214/06-BA107REJ. 5, 6, 12, 17, 26

Vernon, I., Goldstein, M., and Bower, R. G. (2014). "Galaxy Formation: Bayesian History Matching for the Observable Universe." *Statistical Science*, 29(1): 81–90. MR3201849. doi: https://doi.org/10.1214/12-STS412. 17

Vernon, I., Gosling, J. P. (2022). "Supplementary Material for "A Bayesian Computer Model Analysis of Robust Bayesian Analyses"." *Bayesian Analysis.* doi: https://doi.org/10.1214/22-BA1340SUPP. 4

Vicens, G. J., Rodriguez-Iturbe, I., and Schaake, J. C. (1975). "A Bayesian framework for the use of regional information in hydrology." *Water Resources Research*, 11(3): 405–414. 18, 20, 21, 22, 23

Watson, J. and Holmes, C. (2016). "Approximate Models and Robust Decisions." *Statistical Science*, 31(4): 465–489. URL http://dx.doi.org/10.1214/16-STS592 MR3598725. doi: https://doi.org/10.1214/16-STS592. 2

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(7-8): 1703–1729. MR3283898. doi: https://doi.org/10.1137/120900915. 26

Zhu, H., Ibrahim, J., Lee, S., and Zhang, H. (2007). "Perturbation selection and influence measures in local influence analysis." *The Annals of Statistics*, 35(6): 2565–2588. MR2382658. doi: https://doi.org/10.1214/009053607000000343. 2

Acknowledgments

We thank Peter Craig and Richard Wilkinson for valuable discussions when we were researching our method and related topics.