Bayesian Analysis (2010) **5**, Number 4, pp. 697–708

Rejoinder

Ian Vernon[*], Michael Goldstein[†] and Richard G. Bower[‡]

We thank the discussants David Poole, Pritam Ranjan, Earl Lawrence, David Higdon, and David van Dyk for their commentaries on our paper, and for raising many interesting points for discussion, ranging from practical questions related to the implementation of our methodology to fundamental issues of the role and purpose of Bayesian analysis in science. We respond to each of the discussants as follows.

1 Response to David Poole

Poole agrees that often a fully Bayesian analysis for computer models can be difficult and that simplifications such as a the Bayes Linear approach described in our paper are often helpful. He then goes on to discuss another technique known as Bayesian Melding, whereby prior information regarding both the inputs and outputs of the computer model function are amalgamated into a single prior over x. He remarks that such an approach would most likely not be suitable for use on the Galform model due to computational reasons (as melding normally requires complete knowledge of the function $f(x)$). We would state that while it is possible that these computational issues could be resolved by the appropriate use of emulators within the melding calculation, there are unresolved issues about the validity of the melding calculations, consideration of which would have taken us beyond the remit of the study.

Poole then asks "what one loses by doing a Bayes Linear approach?" compared to a hypothetical fully Bayesian analysis. We respond that (as we discuss toward the end of section 3.3), if a fully Bayesian approach were feasible, in that we were prepared to spend the considerable amount of extra time and effort to construct and document realistic joint priors over the input space, the model discrepancy and all other quantities of interest, and if such priors contained extra physical information that was defensible to other cosmology experts in the field, then we would be able to perform a more detailed fully Bayesian inference which would reflect the additional physical information contained within the prior specification.

However, an elicitation for such complex objects would present substantial conceptual and practical difficulty, and hence we perform a Bayes Linear analysis which may be viewed as a pragmatic compromise to such an arduous analysis, as it only requires expert assessments over means and variances. Were we confident enough to obtain more detailed expert judgements, then we can incorporate these within the Bayes Linear framework also (by writing down covariances of more complex objects e.g. higher order quantities (Goldstein and Wooff (2007))). We choose the Bayes Linear method to

[*] Department of Mathematical Sciences, Durham University, Science Laboratories, Durham, UK, mailto:i.r.vernon@durham.ac.uk

[†] Department of Mathematical Sciences, Durham University, Science Laboratories, Durham, UK, mailto:michael.goldstein@durham.ac.uk

[‡] Department of Physics, Durham University, Science Laboratories, Durham, UK, mailto:r.g.bower@durham.ac.uk

DOI:10.1214/10-BA524REJ

obtain meaningful answers in reasonable time using reasonable effort. See the rejoinder to Lawrence and Higdon and to van Dyk for further comparisons between the Bayes Linear and full Bayes approaches.

What we would suggest as unnecessary is the far too common form of Bayesian analysis whereby priors are chosen that have little physical insight and are mainly of forms that provide mathematical convenience for the subsequent challenging fully Bayesian calculations. We have reservations about the value and meaning of such calculations. Were the ensuing calculations simple and transparent, then such an approach, even so, could have considerable exploratory value, but current Bayesian technology does not make this easy.

2 Response to Pritam Ranjan

Ranjan asks about the choices we have made in the history matching process. As described in this paper and summarised in section 3, the history matching approach involves emulating appropriate collections of outputs of the computer model, and combining this with observational data in order to eliminate portions of the input parameter space. Specifically, in this application we chose a subset of 7 outputs of the Galform computer model to compare to the data. This subset increased from 7 to 11 outputs in later waves (these are shown as the vertical lines in figures 12 and 13). Ranjan provides an interesting discussion over this reduction of the data, and asks for a formal technique to pick a subset of outputs that are sufficient to capture most of the characteristics of the functional response of the Galform model. There are techniques for this purpose, and one that we have used in similar applications is Principal Variables (see Cumming and Wooff (2007); Cumming and Goldstein (2009a)).

However, it should be noted that in the first few waves of our study we *do not* seek to identify a fully sufficient set of outputs. We only use a small number of outputs that a) are straightforward to emulate and b) are informative enough to allow us to cut out large portions of the input space (see section 4.2). Once the input space has been reduced, outputs become easier to emulate accurately for reasons discussed in section 3 and demonstrated in section 7: essentially it is because we are 'zooming in' on a locally smooth function. We can then emulate a larger set of outputs to further reduce the input space. This iterative strategy is a major strength of our approach and results directly from our aim of removing implausible (or 'bad') points as opposed to identifying 'good' points. At the end of our analysis we performed a large set of Wave 5 runs to check if the runs that match our 11 outputs also match the other outputs that were not considered. These runs are shown in figures 12 and 13 (bottom right panel) and confirm good matches across all outputs of interest, suggesting that the final set of 11 outputs was indeed sufficient to an acceptable degree.

Ranjan proposes an alternative approach that involves emulating and minimising a function $g(x)$: a global measure that depends on all outputs, which would equal zero if there were an exact match to the data. While this approach may work in some applications, it does have the following disadvantages compared to our method.

Active Inputs: In our approach we initially emulate individual outputs that mainly depend only on a small set of active inputs (e.g. in Wave 1 each output was emulated using sets containing only 5 active inputs) thus greatly simplifying the emulation and analysis. This benefit has been exploited in several applications (see Craig et al. (1997); Cumming and Goldstein (2009a,b)) where the union of active inputs may be large, but the number used for any particular output is small. In contrast, despite $g(x)$ giving scalar output, it will usually depend on all inputs to the function $f(x)$ and hence will be a very complex, high dimensional function. It is usually far easier to emulate a few low dimensional functions than one high dimensional one.

Capturing Physical Dependencies: Emulating individual outputs has a further significant advantage. As discussed in section 3.4, we prefer to build more structure into the mean function or regression part of the emulator for several reasons. Perhaps the most important is that the individual outputs of many physical models, and of Galform in particular, exhibit strong and physically interpretable monotonicities with respect to the inputs, which are naturally expressed through the mean function, resulting in more accurate emulation. The $g(x)$ function would have no such monotonicities, being most likely a complex surface with many local minima, and such emulation advantages would be lost.

Physical Interpretation: There is also the question of the physical interpretation of the emulators that our approach allows: the results of emulation of individual outputs can be checked against expert knowledge and can often help the scientist further understand the behaviour of the model. This again would be lost if we were to analyse only $g(x)$. Our method also allows a more nuanced approach to assessing model adequacy, which we take full advantage of, as we discuss in the response to van Dyk.

Ranjan then suggests recasting our implausibility approach in terms of an expected improvement (EI) criterion. We do fear that the form of an EI algorithm may miss some of the more appealing aspects of our approach, namely that we do not attempt to emulate accurately over the whole input space, only to emulate certain outputs sufficiently accurately to be able to discard large regions of input space. Also, we never express criteria for acceptable inputs (until the Wave 5 runs at the end of the analysis) and only work in terms of non-implausible (i.e. not currently ruled out) inputs. It is unclear if one could devise an EI criterion that incorporates these attributes.

At each wave the current non-implausible volume of input space was checked for connectedness. Ranjan rightly asks whether each of these volumes are also convex in structure. This is an interesting point and we intend to investigate this possibility in future work, as our history matching method could easily produce highly non-convex regions of input space after each wave for certain applications. As concerns Galform, the majority of projections of the non-implausible regions into 2, 3 and 4 dimensional subspaces suggested a convex (or almost convex) shape. The major part of each of our emulators is based on polynomials that can be fitted over non-convex regions (although one has to be careful of the usual traps of fitting over a non-orthogonal design). The Gaussian process part of our emulators have shorter correlation length parameters (as the regression terms take up the global behaviour), and hence we only need to worry about more localised non-convexity, compared to say using a full Gaussian process emulator with simple mean function. These considerations combined with healthy amounts of diagnostics to test each of the emulator's performances (200 diagnostic runs were done at each wave), would suggest that non-convexity is not a significant problem here.

Ranjan's suggested solution, at each wave to use all previous runs from all previous waves to construct an emulator over the whole original input volume, does address the non-convex problem, but for an unacceptable cost. This loses one of the fundamental motivations for our approach which is that it is generally *easier* to emulate the function over smaller volumes. This is discussed in section 3 and demonstrated in section 7 and is mainly due to the polynomial terms becoming better approximations to the smooth function $f(x)$, and the higher density of points allowing the Gaussian process terms to become more accurate.

As to the query raised concerning the practical methodology for constructing the individual emulators, we strongly believe that use of both active variables and a detailed mean function is advantageous in the majority of computer model applications. However, there is flexibility over the choice of techniques used. As we had a reasonably large number of Galform runs, it was felt that traditional model selection techniques were adequate for our purposes. If we had a substantially smaller set of runs then we would have to employ a more careful and formal Bayesian style approach for such selection.

In response to the question about the principles behind our design choices, in order to construct, for example, the Wave 2 design, we decided that computational resources would allow the evaluation of approximately 1400 to 1450 runs. We used a latin hypercube of size 9500 and rejected all the points that did not satisfy our implausibility cutoffs. We chose this size of design as we knew (from experimenting with our emulators and implausibility measures to determine the approximate volume remaining after wave 1) that the number of runs that survive should lie in the range 1400 to 1450. Similar rules were used for subsequent waves.

This leads us to the total number of runs used in this analysis. As this was an application at the cutting edge of investigations into Galaxy formation, we felt it wise to use the substantial computer time available to evaluate as many runs as was feasible (1000+) for each wave. That is, the relatively large number of runs used is a reflection of the computer resources available and the importance of the project. It is clear that we could have performed this analysis with fewer runs, but it was thought best to err on the side of caution. Exactly how many evaluations would have been required to achieve the same goal is an interesting question, and leads to the obvious design issue of how to proportion a fixed quota of runs between various future waves. We intend to look at this important topic in future work.

3 Response to Earl Lawrence and David Higdon

Lawrence and Higdon (LH) requested our Wave 2 data, which was composed of 1414 runs that were restricted to the non-implausible region defined by the Wave 1 implausibility cutoffs (given by equation (22)). We are pleased that they were able to apply their methodology as described in Higdon et al. (2008) to this data set, and that they have produced some interesting, if provisional, results in time for this discussion.

It seems that the fully Bayesian approach is more "aggressive" in reducing input space than the Bayes linear version, but it would at this stage appear to be too aggressive. Comparison with figure 11 which shows a marginal plot of the set of Wave 5 runs, coloured by implausibility, shows several acceptable (green) runs that are clearly outside of the hpd region given by LH's analysis (see for example the $V_{\rm hotdisk} : \alpha_{\rm hot}$, the $\alpha_{\rm cool} : \alpha_{\rm hot}$, the $\epsilon_\star^{-1} : p_{\rm yield}$ and the $V_{\rm hotdisk} : \alpha_{\rm reheat}$ projections). That is to say, some of the acceptable runs which are of interest to the cosmologists (the corresponding luminosity functions of which are shown in the bottom-right panel of figures 12 and 13) could be excluded by this analysis.

Lawrence and Higdon urge caution in making direct comparisons between the two approaches based on the marginal posterior densities given in figure 1 of their discussion paper, as the two cases are not equivalent. For example a major difference is that the posterior as calculated by LH used the model discrepancy Φ_E conditioned on a fixed value of the parameters a, b and c (see equation (20)). The implausibility approach explored the full ranges of the parameters a, b and c as specified by the expert and given by equation (21), by only discarding an input x as implausible if it failed the implausibility cutoffs for all values of a, b and c. Incorporating the uncertainty on the parameters a, b and c into LH's approach could possibly lead to a more diffuse posterior and make a comparison between the approaches easier (and may go someway toward solving the Wave 5 run problem outlined above). As LH state, the differences in the emulation approaches used will also muddy this comparison.

A general problem is that is it not at all obvious that the posterior shown in LH's figure 1 (which was generated using the Wave 2 runs) still respects the Wave 1 constraints that make the analysis meaningful. That is to say, are there parts of LH's posterior that give significant probability to inputs that were previously ruled out by the Wave 1 constraint? LH's assessment can only be based on an extrapolation from a subspace covered by the wave 2 runs whereas the wave 1 elimination was based on function evaluations more local to that region. There is no guarantee that this issue is avoided simply because the 2-dimension marginals of LH's figure 1 are seen to be within the 2-dimensional projected non-implausible region of figure 10, say.

LH then go on to describe their impressive work on the Coyote Universe (Heitmann et al. (2009)), where, as they state, the main computational effort went into finding suitably smooth representations of the power law (see LH's figure 4), while retaining the vital physical features (known as the baryonic acoustic oscillations), important for understanding structure formation. They use PCA to capture the behaviour of all outputs of the computer model. PCA is widely used in the computer model literature, and has many obvious benefits. However, it has some disadvantages too. Often the most important principal components will depend on all of the active inputs to the computer model and hence emulating them may be difficult. This should be compared with the Principal Variables approach discussed in the response to Ranjan above, where each principal variable may only depend on a small subset of the active inputs. Principal variables have similar properties to principal components in their ability to largely reconstruct the entire output set.

Although our approach differs from LH's due to the use of Bayes linear methods as opposed to a more fully Bayesian treatment, another fundamental difference is that between history matching and calibration. LH are performing calibration and hence assume there exists a single input (the "best input" x^+, as defined in section 3), and subsequently try to calculate the posterior distribution for the single point x^+. In the history matching approach we ask the more general question: which inputs x are not obviously inconsistent with the notion of x^+? It is not surprising that these approaches give different but related answers. In particular the statement that there is a unique best input may often sharpen the credible intervals, as compared to a Bayes linear or even fully Bayesian history match. We discuss this difference in more detail in the response to van Dyk below.

We highlight the above differences between their approach and that of our own to emphasize the many subtleties involved in such computer model analysis. We would recommend to anyone attempting a fully Bayesian calibration of such a complex computer model, to consider preceding it by a history match. (Our understanding is that this is why LH asked for the wave 2 runs rather than those of wave 1). A history match should be performed to identify if there are any acceptable matches and their location in input space (which is often of great interest to the modeller), and then a fully Bayesian calibration should be performed only over the restricted input region defined by the history match. Performing an MCMC algorithm whilst respecting various complex constraints as provided by the history match may present some interesting challenges: see our concerns about this above and in the response to van Dyk.

This combined process is equivalent to cutting out the often large regions of input space that would have extremely low posterior probability, before proceeding with the Bayesian analysis, and should result in a highly accurate approximation to the posterior distribution. We intend to explore this, powerful strategy, in future work.

4 Response to David van Dyk

We thank David van Dyk for his comments, and for the opportunity to expand further on some aspects of our work. His comments raise interesting questions about the meaning, the potential and the limitations of Bayesian investigations within fundamental science. We deal with his points in turn.

The first point raised is a query about quality of fit in figure 14. This figure (which shows new types of outputs not considered in this work) is included purely to show the next stage of the matching process. As we have now identified the region of input space consistent with both the bj and K luminosity function observed data, we are now free to move around this region, exploring the effect on the new outputs shown in figure 14. In this way, we view history matching as an ongoing process, notably simpler than attempting to incorporate all data constraints simultaneously.

Van Dyk proceeds to give a description of history matching in terms of "standard statistical methodology" using the log-likelihood $L(\theta|Y)$. While there are similarities between this description and our methods, van Dyk seriously oversimplifies some crucial features.

θ **does not exist:** Most importantly, we do not assume that there exists a single "true" input θ (or best input x^+), as even though the inputs to the Galform model are related to real physical quantities, they are not themselves physical. Van Dyk's summary assumes that θ exists as a true but unknown quantity. However, the actual situation is that θ is largely a model construct and that as the Galform model evolves over future generations, the interpretation of the various elements of θ will change, with some even ceasing to exist. As mentioned in the response to Lawrence and Higdon, not assuming the existence of θ changes the questions that one might ask. We ask only if there are any values of the inputs x that are not inconsistent with the concept of such a best input θ. We feel that it is essential to establish the answer to this question, before considering whether a calibration is appropriate.

Implausibility Measure is not a log-likelihood: Even ignoring this issue, the log-likelihood that van Dyk writes down is not comparable to the problem we analyse as he has ignored the θ dependance of σ_i, which comes from the use of emulators and is of great importance in this context. That is to say, a more appropriate comparison would be if $Y_i \sim \mathrm{N}(\mu_i(\theta), \sigma_i^2(\theta))$, where $\mu_i(\theta)$ and $\sigma_i(\theta)$ both depend on θ. In this case the log-likelihood is now, ignoring constants, $L(\theta|Y) = -\frac{1}{2}\log|\Sigma(\theta)| - \sum_{i=1}^{n}(Y_i - \mu_i(\theta))^2/2\sigma_i^2(\theta)$, where in this simple illustration the matrix $\Sigma(\theta) = \mathrm{diag}\,\sigma_1^2, .., \sigma_n^2$. Thus $L(\theta|Y)$ contains an additional $-\frac{1}{2}\log|\Sigma(\theta)|$ term which *does not feature* in our implausibility measures, as this term comes directly from the full distributional assumption of normality, which we do not make.

In summary our approach is both mathematically distinct and different in fundamental interpretation from the interpretation van Dyk suggests, and from most of the approaches used in the literature. As discussed below, much of the computer model terminology is used to highlight these essential differences.

In the section entitled "**Employing the Common Statistical Framework for Computer Modelling**", it is remarked that "(the authors) aim to find the values of the parameter that result in the best fit to the data". We in fact aim for the opposite: to find the input parameters x that are clearly not good fits to the considered data, given current knowledge of the computer simulator $f(x)$. We then discard these inputs, leaving a hopefully non-empty set of input parameters that are deemed non-implausible. As we perform more runs of the computer model, or bring into consideration more outputs, this set will decrease in size. At no point are we trying to find the "best" set of input parameters, just those that give acceptable matches.

This essential difference between discarding 'bad' inputs and searching for 'good' inputs is critical, coming directly from the lack of a best input assumption, and is the reason for the power of our sequential approach. It is far easier to emulate a small set of outputs of the function in order to determine that large parts of the input space are implausible (and continue this process iteratively), than it is to attempt to identify the “best fits” which would require emulation of all outputs, use of all observed data and careful modelling of every part of the problem. Even if we are searching for good matches, it is a sensible way to simplify massively the calculations by first eliminating bad matches.

Often the different terminology in use in the computer model literature has arisen due to the subtly different problems presented in this area. For example, history matching, the process of iteratively discarding implausible inputs as outlined above, is not equivalent to ‘model checking’ which refers to confirming all aspect of the stochastic formulation. History matching is a process applied to the computer model (i.e. the function $f(x)$) itself. The term ‘calibration’ is a statistical term widely used in inverse regression problems, of which the calibration of a computer model through use of an emulator can be viewed as a direct generalisation.

As discussed above, an implausibility measure is *not* the same as the likelihood. The former is only informative regarding unacceptable (i.e. implausible) inputs and says nothing about possible good inputs (a low implausible value $I(x)$ should be interpreted as “x is not ruled out yet”): it is deliberately not normed. The likelihood comes from carefully modelling all aspects of the data (a sometimes arduous task in computer model problems), and is informative regarding both ‘good’ and ‘bad’ inputs. The likelihood in computer model applications is unfortunately often non-robust. We are somewhat surprised as to van Dyk’s comments regarding the use of the term “inputs”. We use the term in the mathematical sense to refer to the inputs x to the Galform function $f(x)$, and initially introduce them in the earlier sections as “input parameters” to avoid confusion (see for example the abstract, section 1, section 2 and specifically table 1). Thus they are distinguished from the other parameters in the problem e.g. those used in fitting the emulators: β_{ij}, σ_i, ω_i and θ_i.

This is not just a case of different terminology, but rather a case of labelling the specific types of problems faced in a computer model analysis that are often quite different from those encountered in other statistical applications. Indeed, it is possible to go further and remark that (as van Dyk does point out), statisticians can learn from some of the ideas presented in the computer model literature. Specifically, the reasoning leading to the clear and upfront acknowledgement that the computer model is not a perfect representation of reality could also be applied to any statistical model too (as is discussed in Goldstein (2010)). As van Dyk remarks, such computer models are often embedded within large statistical models to enable analysis of complex systems. Care must be taken in such situations not to oversimplify: the danger is that the computer model is treated as reality, and no model discrepancy is used.

Van Dyk next argues that the multivariate implausibility measure $I(x)$ given by equation (16) should be considered superior to $I_M(x)$ (and implicitly $I_{2M}(x)$ and $I_{3M}(x)$) which are the first, second and third highest univariate implausibilities corresponding to each individual output, respectively, and given by equations (13), (14) and (15). Again, as we are not attempting to identify likely values of the parameters, we would argue that the choice between univariate and multivariate measures is highly problem dependent, and that in many cases one can pay too high a price for using the multivariate measure

$I(x)$ too early. This measure is very sensitive to possible failings of the emulators (it generally requires the construction of an accurate multivariate emulator, which is often difficult). It also requires a full multivariate model discrepancy specification which can be both hard to elicit and to document (see for example equations (20) and (21) and the accompanying discussions). For these reasons, we recommend using the conceptually simpler, easier to elicit, and more robust measures $I_{2M}(x)$ and $I_{3M}(x)$ to reduce the input parameter space in the initial waves. Then, as we have done in the current work, $I(x)$ can be brought in for use at a later wave, when emulation is easier. It would of course be interesting, if possible, to track the changes in $I(x)$ as one progresses from wave 1 onward.

In the section "**What does it mean to be a Bayesian**" van Dyk gives some interesting views about the Bayesian paradigm, many of which we agree with. Having said this, the statement "the Bayes Linear approach is based on Bayes Theorem" is misleading: as is discussed at the end of section 3.3, the Bayes Linear approach is a generalisation of Bayes Theorem, based on taking expectation rather than probability as the natural primitive for the subjectivist theory (see De Finetti (1974) and Goldstein (2006) for more details). We would agree that one of the major benefits of a Bayesian analysis is that "it is a principled analysis that fully accounts for the complexities of the underlying distributions and avoids the old and often unrealistic Gaussian assumptions", but only provided serious effort is put into capturing the beliefs of the expert, and into creating a realistic statistical model, so that the underlying distributions used have actual physical meaning, as opposed to being based on arbitrary assumptions or mathematical simplicity. Our perspective is that often this level of detailed elicitation is infeasible, in which case we would rather make a simpler specification based on means and variances and proceed with a tractable Bayes Linear analysis, than make arbitrary assumptions as to the forms of several distributions and try to proceed with an often difficult and (in the case of computer models) highly non-robust Bayesian analysis.

We now turn to the discussion of expert judgement within our analysis. First it should be noted that the judgements asserting that the quantities in equations (1) and (2) are uncorrelated actually represented the assessments made by the expert after extensive discussion and consideration. We also note that similar assumptions are standard throughout the field of computer models and used in countless papers (e.g Kennedy and O'Hagan (2001)), often with a much stronger assumption of full independence. It is possible to consider the structural beliefs about model discrepancy even more carefully (as is discussed in Goldstein and Rougier (2009)), but we considered the current assessment adequate for the purposes of this study.

The most significant area of our analysis which involves expert judgement is that of the model discrepancy ϵ_{md} which represents the difference between the model output and reality itself. This was broken down into three contributions, two of which were assessed using further computer model runs, while the third component Φ_E came directly from Richard Bower's expert judgements. In a Bayesian analysis of any form, it is impossible to address the issue of model discrepancy meaningfully without using expert judgements of this kind. Crucially, in this project we had the benefit of prolonged contact with the expert resulting from regular monthly meetings for over two years. The judgements themselves were formed from many considerations of the possible deficiencies of the model (missing physics, approximate models of certain processes, inaccurate dark matter simulations) only some of which we were able to discuss in the paper due to length restrictions. The judgements were made by an expert fully versed in the meaning and impact of Φ_E. In this work, the model discrepancy should be understood in terms of the expert's tolerance for what would be classified as an acceptable run, and therefore the model discrepancy is no longer an uncertain quantity to be estimated using standard statistical methods.

The assessment as to whether the Galform model is an adequate model of Galaxy Formation *can only come from expert judgement.* Van Dyk is clearly concerned that "the expert assessments determine the final outcome". This misunderstanding arises due to the belief that the Galform model can be deemed acceptable by some abstract notion of 'right' or 'wrong'. This is an assertion that is disconnected from the way in which scientists view models of this complexity. Scientists are always concerned both with incremental and paradigmatic improvements to their model. If, for example, we found no acceptable runs, we would increase the tolerance as represented by the model discrepancy, to determine how large it would have to be to obtain some 'acceptable' runs. This would be informative for the scientists as it would give some measure of how inadequate the model is, and show the impact of the missing (or incorrect) physics in the current model formulation. Such measures are crucial in helping scientists to assess whether incremental modifications will be adequate to deal with the observed discrepancy between the model output and physical observations. Note that, our approach explicitly incorporates a sensitivity analysis as relates to the expert assessment of Φ_E. As can be seen in equations (20) and (21) the $\mathrm{Var}(\Phi_E)$ is parameterised by three parameters a, b and c that the expert was unwilling to assign specific values. Therefore we explored the impact on our implausibility measures of varying the parameters a, b and c over ranges agreed by the expert and given by equation (21). We only discarded an input x as implausible if it failed the implausibility cutoffs for *every* value of a, b and c in the given ranges. We also performed sensitivity analysis on the choice of implausibility cutoffs imposed using various sets of diagnostic runs. A benefit of our approach is that such sensitivity analysis can be relatively straightforward.

Van Dyk states "in my view even a Bayesian analysis must work hard to minimise its assumptions and must be absolutely upfront about the impact of its subjective assessments on the final analysis." We could not agree with this statement more, provided we are clear about the distinction between assumptions which are by necessity somewhat arbitrary, and subjective assessments, which reflect the careful and knowledgeable assertions of experts. Rather than making several unjustifiable distributional assumptions (which are necessarily somewhat arbitrary and correspond to an infinite number of probabilistic assertions) whose impact on the posterior is in many cases extremely hard to assess, there often are advantages in using a Bayes linear style approach involving only a relatively small number of first and second order quantities, where subjective assessments on quantities such as the model discrepancy enter clearly, and the effects of which are easy to demonstrate. Having said all this, we would be interested to see a full elicitation as applied to Galform, followed by a Bayesian inference in which each distributional statement is robustly defended.

In the section entitled "**If it looks like a Duck...**" the implausibility measures are again equated to the Gaussian log-likelihood, which as discussed above, is an inappropriate comparison. For example, if we had judged Gaussian distributional assumptions to be appropriate, we would have used much tighter credible volumes and cutoffs as we would not have had to appeal to Pukelsheim's 3σ rule (the very powerful result that for any unimodal, continuous distribution more than 95% of its probability lies within $\pm 3\sigma$) and instead used a more familiar 2σ rule (as incidentally it appears LH have done). Such differences are always present when comparing a Bayes Linear approach to a full Gaussian specification in any example.

Van Dyk then asks directly about the differences between the two approaches in this application: they are substantial as we now describe. A fully Gaussian Bayesian analysis would attempt to construct a likelihood based on all available data points and an assertion of a unique true value for θ. It would further depend on constructing multivariate emulators to represent jointly every single output of the computer model for which there is data, even if some of these outputs are very difficult to emulate accurately over the whole input space. The modelling assumptions and emulator construction that would go into building this multimodal likelihood would need to be highly accurate, otherwise it would result in any inference being extremely non-robust. The approach we describe in this work avoids both the assertion of the unique true value of the parameters and much of the difficulty in the modelling and analysis. We deal with implausibility measures that are far more robust being functions of small numbers of outputs (outputs which are chosen for their ease of emulation). At each wave we do not need to model any highly multimodal likelihood: all we need to do is draw a conservative contour around the low implausibility input points. As demonstrated in section 7, at each wave the emulators improve in accuracy and we can consider more output data constraints when necessary. Note that as we introduce the constraints from the data sequentially, even a large number of constraints is often straightforward to incorporate, unlike for an MCMC algorithm for which it may be problematic.

In "**The scientific objective**" van Dyk asks some questions about the goal of this analysis. The appropriateness of the analysis depends upon the context of the scientific question. If we are looking at well defined physical quantities (location, age, metal content) then often estimating these quantities would be the appropriate choice of analysis. However the Galform input parameters are more abstract, and only make sense in terms of the model and its applicability. In this case we are trying to assess *whether* Galform can shed light onto the physical world, or at least contribute to that large question. An approach where we perform a Bayesian calibration for the best input x^+, may become appropriate when the model has been shown to be sufficiently accurate for all intended purposes. Provided that the history match has revealed the model does appear to be sufficiently accurate, the discussant's suggestions in the section "**The Final Analysis**" may be helpful for achieving such a Bayesian calibration. As a general principal, carrying out a history match as a first stage in the Bayesian calibration exercise will usually be very useful in greatly restricting the volume of parameter space that needs to be explored by the MCMC algorithm.

References

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." 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: Springer-Verlag. 699

Cumming, J. A. and Goldstein, M. (2009a). "Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments." In O'Hagan, A. and West, M. (eds.), *Handbook of Bayesian Analysis*. Oxford, UK: Oxford University Press. 698, 699

— (2009b). "Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations." *Technometrics*, 51(4): 366–376. 699

Cumming, J. A. and Wooff, D. A. (2007). "Dimension reduction via principal variables." *Computational Statistics & Data Analysis*, 52(3): 550–565. 698

De Finetti, B. (1974). *Theory of Probability*, volume 1. London: Wiley. 705

Goldstein, M. (2006). "Subjective Bayesian Analysis: Principles and Practice." *Bayesian Analysis*, 1(3): 403–420. 705

— (2010). "External Bayesian analysis for computer simulators." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *To appear in Bayesian Statistics 9*. Oxford University Press. 704

Goldstein, M. and Rougier, J. C. (2009). "Reified Bayesian modelling and inference for physical systems (with Discussion)." *Journal of Statistical Planning and Inference*, 139(3): 1221–1239. 705

Goldstein, M. and Wooff, D. A. (2007). *Bayes Linear Statistics: Theory and Methods*. Chichester: Wiley. 697

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. 701

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). "Computer Model Calibration Using High-Dimensional Output." *Journal of the American Statistical Association*, 103(482): 570–583. 701

Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society, Series B*, 63(3): 425–464. 705