# Exchangeable Computer Models

Leanna House

*Virginia Tech, Blacksburg, VA, USA.*

Michael Goldstein

*Durham University, Durham, UK.*

Ian Vernon

*Durham University, Durham, UK.*

**Summary**. We address the uncertainty of deterministic computer models that rely on both input parameters and system conditions. We refer to such models as multi-deterministic. Multi-deterministic models allow the system condition to vary, and thus have the potential to produce more than one result per input. We introduce the notion of latent computer model outcomes which respond to the results of the multi-deterministic model when using the appropriate, but unknown, system condition for the physical system of interest. The goal for this paper is to make inferences about the latent model given a sequence of realized multi-deterministic model evaluations. We consider the case where the sequence of models is judged a priori to be second order exchangeable over the system condition and use Bayes linear methods to assess the posterior expectation and variance of the latent model given the realised evaluations. We demonstrate our methods using multi-deterministic results from a galaxy formation model called Galform for which the system condition is the specfication of dark matter over time and space.

## 1. Introduction

Many natural phenomena, such as, volcano eruptions, daily temperature, and star formation, are influenced by underlying physical processes which are not fully understood. To learn about such processes, observational data is often supplemented by synthetic datasets manufactured by mathematical models which characterize the processes theoretically and are embedded in computer simulation experiments. However, with the benefits of using synthetic data, comes the responsibility of accurately assessing the uncertainty of the mathematical model. In this paper, we develop a novel way to assess one source of uncertainty for a type of model which we refer to as multi-deterministic. We consider a model to be "multi-", not purely, deterministic when a range of output values from a computer model may result deterministically for the same input parameter values, provided differing system conditions. An assessment of the variation in output due to multiple conditions is vital for understanding the uncertainty of model based inferences. The goal of this paper is to show that we can make the required assessment of uncertainty, if we consider the versions of the computer model evaluated under different system condition choices to be a sequence of second-order exchangeable computer models.

A multi-deterministic model can be considered to be a collection of deterministic models which share the same domain, share the same co-domain, do not present obvious theoretic or computational advantages, and generate disparate predictions. In the presence of multiple models, current approaches tend to base model predictions on either the outcomes from

one (arguably, the "best") model or the outcomes averaged, per input, across conditions. However, such methods will likely underestimate the prediction uncertainty. For example, consider an unobserved or latent model that corresponds to the physical process in the sense that the model actually relies on the appropriate, or possibly true, system condition. Methods that average one or more realized computer model rely on the strong assumption that the average equals the latent model without error. For this paper, we choose not to ignore the possibility of error and propose using the outcomes from multiple models to estimate the latent model with uncertainty. In turn, inferences about the true process will be based on our latent model assessments.

Many computer models that are treated currently as deterministic, should be considered as multi-deterministic and have the potential to estimate a latent model. In Section 2, we provide a non-exhaustive list of multi-deterministic models which incorporate system conditions differently. In particular, we describe a model called Galform which we use in Section 9 to demonstrate our methods. Galform simulates the formation of one million galaxies within our universe and includes a system condition that represents dark matter. The exact configuration of dark matter present throughout the evolution of our universe is unknown, so any specification of dark matter will be uncertain and the choice will affect model outcomes. From approximately 1000 model evaluations for 40 different dark matter specifications, our methods enable us to construct an emulator for the latent Galform outcomes, i.e., outcomes based upon the dark matter configuration which actually underlies our universe. For our analysis, we consider the outcomes from Galform per system condition specification to be second-order exchangeable with both each other and the latent outcomes, and we use Bayes linear methods to make posterior moment assessments.

The outline of the paper is as follows. In Section 2, we describe several common multi-deterministic models and establish notation. In Section 3, we define the notion of a latent model and state explicitly the analysis goals for this paper. In Section 4 we discuss the second-order exchangeability assumptions that we make to analyze a latent model given multi-deterministic model results. We explain the role of emulators in our analysis and how to estimate them in Sections 5 and 6. We then develop the theory for making Bayes linear adjustments to the estimated emulators in Section 7 given multi-deterministic data. In Section 8, we describe how to elicit prior judgments about model parameters from experts, and we apply our methods to Galform in Section 9. We conclude with a discussion of our work in Section 10.

## 2. Multi-deterministic Models

Any deterministic model that relies on a pre-specified system condition, such as a set of initial conditions or a forcing function, may be viewed as a multi-deterministic model. Consider a deterministic computer model $f$ that depends upon both a system condition $c$ and input vector $x$ ($x = [x_1, ..., x_p]$) and returns $f^{(c)}(x)$ ($f^{(c)}(x) = [f_1^{(c)}(x), .., f_q^{(c)}(x)]$). Purely deterministic models fix the condition, e.g., $c = c_j$, so that for $x = x_i$ and $x = x_{i'}$, $f^{(c_j)}(x_i) = f^{(c_j)}(x_{i'})$ when $x_i = x_{i'}$. Multi-deterministic models, however, allow for model results to differ for equal inputs because specifications for $c$ may vary; if $c_j \neq c_{j'}$, $f^{(c_j)}(x_i)$ need not equal $f^{(c_{j'})}(x_i)$, even though the inputs $x_i$ are the same. In this section, we provide common examples of multi-deterministic models and define the notation we use for the remainder of the paper.

### *2.1. Examples*

For some multi-deterministic models, the purpose for the system condition in the computer model is indistinguishable from the role of any other input variable. In such cases, the dependence structure of $f^{(c)}(x)$ on $c$ might have a reasonable parametric form and a statistical analysis for the computer model may simply treat the system condition as an additional input variable. However, this paper addresses the uncertainty of model based inferences when the relationship between the system condition $c$ and the functional output $f^{(c)}(x)$ is not of direct interest, but an assessment of the variation in model predictions due to $c$ is still important. If we were to compare these multi-deterministic models to a mixed statistical model (McCulloch, 2002) which contains both fixed and random effects, the system condition might correspond to the random effect. The following describes four examples of multi-deterministic models.

(a) The quantity $c$ may represent an unknown physical condition that is too complex to be assessed parametrically. For example, some scientific theories suggest that the processes controlling the evolution of our observable universe rely on the presence of dark matter. However, the actual dark matter present since the Big Bang is currently unmeasurable, so any galaxy formation computer model (e.g, Galform, Section 9) that relies on the specification of dark matter is inherently uncertain. Similar examples include $c$ representing computer model forcings, such as boundary conditions, continental configurations, global elevations, and a vegetation distribution, fixed ocean currents, solar variation etc. (e.g., Sewall et al., 2007). Also, $c$ may associate with initial condition specifications for models spun up to equilibrium.

(b) The quantity $c$ may represent a condition that is only measurable with some degree of variance or measurement error. System condition specifications based on observational data result in a range of possibilities for $c$, hence a range for model output, that is based on the limitations in data collection technologies. For example, ocean current models that include land configurations or rainfall-runoff models based on precipitation measurements induce uncertainty in $f(x)$ that is unrelated to the interactions of the process variables $x$.

(c) The quantity $c$ may index model outcomes which relied on differing known conditions that are, in practice, unquantifiably different. For example, climate or crop-yield models may simulate the effects of environmental changes in different global regions. Depending upon the model resolution, each region might contain mountains, bodies of water, vegetation, animal life, urban developments etc. that collectively affect model output, but are, in practice, too difficult to quantify. The variation due to region is not practical to characterize and the region serves as the experimental unit for the computer experiment.

(d) The quantity $c$ may represent a set of model results that condition on fixing a subset of input variables to expert elicited values because they were identified as inactive or due to cost constraints. For example, if $x = [x_a, x_b]$ a researcher might fix $x_a$ to reasonable values in hopes of exploring the space of $f(x_b)$ thoroughly. In some cases however, setting a subset of inputs to constant values may limit the outcome range to unknown regions of the output space, and adjustments should be made to account for this bias.

### 2.2. Notation

We can formulate all of the above examples with the following notation. The condition and inputs, $c$ and $x$, may take values $c_j$ and $x_i$ ($x_i = [x_{i1}, .., x_{ip}]$) respectively, where $c_j \in \mathcal{C}$ and $x_i \in \mathcal{X}$. For a computer experiment which explores $m$ conditions and $n^{(c)}$ values for $x$, let $m^{(x_i)}$, $n^{(c_j)}$, $m$, and $n$ represent respectively the number of conditions explored for $x = x_i$, the number of runs completed for condition $c_j$, the number of different conditions explored, and the total number of runs across conditions, $n = \sum_{j=1}^{m} n^{(c_j)}$. We denote the outcomes from an experiment as sequences where $\{f(x)\}_{[m^{(x)}]}$ contains the model outcomes across conditions for input $x$; $\{f^c\}_{[n^{(c)}]}$ contains the model results for condition $c$ across inputs; and $\{f\}_{[n]}$ contains all of the results, across conditions and inputs.

The models and methods discussed in this paper apply to any computer experiment with an unknown system condition. However, balanced designs for multi-deterministic models have certain powerful theoretical properties which are very useful for analyzing model uncertainty. Therefore, to highlight these properties and to simplify our account, we develop the statistical theory for balanced computer experiments. Such experiments call for the same experimental design per system condition specification so that 1) the same collection of model evaluations, $x_1, ..., x_{n^{(c)}}$ is made for each selected condition $c_j$, and 2) $n^{(c_j)} = n^{(c_{j'})} = n^{(c)} > p$ for all $j$, and 3) $m^{(x_i)} = m^{(x_{i'})} = m$. When necessary, we discuss theoretical issues for unbalanced experiments.

## 3. Goal: Assess Latent Model Outcomes

Multi-deterministic models are useful because researchers may explore multiple specifications for a system condition, when the ideal or true value for the condition is unknown or impossible to specify. We refer to the computer model given the ideal condition as the latent model and denote the outcomes by $f^{(L)}(x)$. For example, the hypothetical or latent model for Galform (Example 2.1.a and Section 9) is based on the true dark matter configuration that existed and evolved over the last 13.7 billion years (the time elapsed since the big bang). In this paper, we develop the methodology to assess the expectation and variance of $f^{(L)}(x)$ for all $x$ provided outcomes from realized versions of the model $\{f\}_{[n]}$.

To model $f^{(L)}(x)$, we rely on the true condition being a priori second-order exchangeable (SOE) with the set of specified system conditions. Second-order exchangeability dictates a fundamental correlation structure between the latent and realized model outputs so that we may use information in the realized outcomes to assess $f^{(L)}(x)$. In the next section, we explain the SOE assumption and extend it to apply to functional data.

## 4. Second Order Exchangeable Functions

As described by de Finetti (1974), an infinite sequence is fully exchangeable when any subset of size $k$ ($k < \infty$) elements has the same joint probability distribution as any other $k$ subset. Thus, the specification of this distribution for each $k$ is required to assess full exchangeability and implies that the probability for an infinite number of finite sets of random variables can be specified a priori. A second-order exchangeable (SOE) sequence however, does not have the same, strong requirement and relies only on judgments for every pair of elements (k=2) in a sequence. In this section, we extend the notion of SOE to functions and show how it applies to multi-deterministic models.

### *4.1. Extending SOE to SOEF*

An infinite sequence of random vectors, say $a_1, a_2, ...$, is SOE when the first and second-order belief specification for the sequence of vectors is unaffected by any permutation of the order of the vectors (Goldstein and Woof, 2007, pg. 184). Thus, for any $j$ and $k$, the following expectation, variance, and covariance are constant,

$$\mathrm{E}[a_j] \;=\; \mu_a, \qquad \mathrm{Var}[a_j] = \Sigma_a, \qquad \mathrm{Cov}[a_j, a_k] = \Gamma_a \;\; \forall\, j \neq k. \tag{1}$$

We introduce the notion of second-order exchangeable functions (SOEF) and state the SOEF Representation Theorem by extending the definition of SOE sequences.

**Second-order Exchangeable Functions**

Let $a(x)$ represent function $a$ with input $x$ where $x = [x_1, ..., x_p]$ and $a(x) = [a_1(x), .., a_q(x)]$. The sequence of functions $\{a^{(1)}, a^{(2)}, ...\}$ is *second-order exchangeable* if,

(a) For the same input $x$, the elements of the sequence $\{a^{(1)}(x), a^{(2)}(x), ...\}$ are SOE with mean $\mu_{a(x)}$ and variance matrix $\Sigma_{a(x)}$

(b) For differing inputs $x$ and $x'$, any two different outcomes have a constant covariance within and between models,

$$\mathrm{Cov}[a^{(j)}(x), a^{(j')}(x')] \;=\; \begin{cases} \Gamma_{a;x;x'} & \text{for } j = j' \\ \Lambda_{a;x;x'} & \text{for } \; j \neq j' \end{cases} \tag{2}$$

and for $x = x'$, $\Gamma_{a;x;x'} = \Sigma_{a(x)}$ .

We have the following representation theorem for SOEF which generalizes the corresponding representation theorem for SOE sequences.

**SOEF Representation Theorem**

Let $a^{(i)}$ represent the $i$th deterministic function in the infinite sequence $\{a\} = \{a^{(1)}, a^{(2)}, \ldots\}$. If $\{a\}$ is SOEF then we may introduce the further random function $\mathcal{M}\{a\}$, termed the population mean function, and also the infinite sequence, $\mathcal{R}_1\{a\}, \mathcal{R}_2\{a\}, \ldots$ termed the individual residual functions which satisfy the following properties:

(a) Given $x \in \mathcal{X}$, for each $j$

$$a^{(j)}(x) = \mathcal{M}\{a(x)\} + \mathcal{R}_j\{a(x)\}$$

(b) The first two moments of $\mathcal{M}\{a(x)\}$ are as follows

$$\mathrm{E}[\mathcal{M}\{a(x)\}] = \mu_{a(x)}, \qquad \mathrm{Var}[\mathcal{M}\{a(x)\}] = \Sigma_{\mathcal{M}\{a(x)\}} = \Lambda_{a;x;x}$$
$$\mathrm{Cov}[\mathcal{M}\{a(x)\}, \mathcal{M}\{a(x')\}] = \Lambda_{a;x;x'}.$$

(c) The sequence $\{\mathcal{R}_1\{a\}, \mathcal{R}_2\{a\}, \ldots\}$ is SOEF, where for any $j, j' \in \mathbb{N}$, $j \neq j'$, $x, x' \in \mathcal{X}$, and

$$\begin{aligned} &\mathrm{E}[\mathcal{R}_j\{a(x)\}] = 0, \\ &\mathrm{Var}[\mathcal{R}_j\{a(x)\}] = \Sigma_{a(x)} - \Sigma_{\mathcal{M}\{a(x)\}} = \Sigma_{\mathcal{R}\{a(x)\}}, \\ &\mathrm{Cov}[\mathcal{R}_j\{a(x)\}, \mathcal{M}\{a(x')\}] = \mathrm{Cov}[\mathcal{R}_j\{a(x)\}, \mathcal{R}_{j'}\{a(x')\}] = 0, \\ &\mathrm{Cov}[\mathcal{R}_j\{a(x)\}, \mathcal{R}_j\{a(x')\}] = \Gamma_{a;x;x'} - \Lambda_{a;x;x'}. \end{aligned}$$

The proof for the SOEF Representation Theorem follows directly from the proof of the corresponding representation theorem for SOE sequences in (Goldstein, 1986; Goldstein and Woof, 2007). For any infinite SOE sequence $\{a_1, a_2, \ldots\}$, each $a_i$ can be written as $a_i = \mathcal{M}\{a\} + \mathcal{R}_i\{a\}$ where $\{\mathcal{R}_1\{a\}, \mathcal{R}_2\{a\}, \ldots\}$ is a mean zero, SOE sequence of mutually uncorrelated quantities and each element is uncorrelated with $\mathcal{M}\{a\}$; and for any two infinite SOE sequences, $\{a_1, a_2, \ldots\}$ and $\{b_1, b_2, \ldots\}$ where the $\mathrm{Cov}[a_j, b_{j'}]$ is constant for $j \neq j'$, the $\mathrm{Cov}[\mathcal{R}_j\{a\}, \mathcal{R}_{j'}\{b\}] = 0$. Properties a-c follow accordingly.

### 4.2. SOEF Representations for Multi-deterministic Models

Within the context of multi-deterministic models, there is in principle an infinite sequence of conditions $\{c_1, c_2, ...\}$ for which, in many cases, we may make the judgment that the functions $\{f^{(c_1)}, f^{(c_2)}, ...\}$ are SOEF. If we make the additional judgment that $f^{(L)}$ is second-order exchangeable with $f^{(c_j)}$ for each $c_j \in \mathcal{C}$ (in the sense that the augmented sequence $\{f^{(L)}, f^{(c_1)}, f^{(c_2)}, ...\}$ is also SOEF), we may consider the SOEF Representation theorem and characterize the latent model as

$$f^{(L)}(x) = \mathcal{M}\{f(x)\} + \mathcal{R}_L\{f(x)\}, \tag{3}$$

which formalizes the relationship between the latent model and any realized set of model outcomes as, for each $x$, $x'$ and $j \neq j'$, we have

$$\begin{aligned}
\mathrm{E}[f^{(L)}(x)] &= \mathrm{E}[f^{(c_j)}(x)] = \mu_{f(x)} && (4)\\
\mathrm{Var}[f^{(L)}(x)] &= \mathrm{Var}[f^{(c_j)}(x)] = \Sigma_{\mathcal{M}\{f(x)\}} + \Sigma_{\mathcal{R}\{f(x)\}} = \Sigma_{f(x)} && (5)\\
\mathrm{Cov}[f^{(L)}(x), f^{(c_j)}(x')] &= \mathrm{Cov}[f^{(c_j)}(x), f^{(c_{j'})}(x')] = \Lambda_{f;x;x'}. && (6)
\end{aligned}$$

We assess $\mu_{f(x)}$, $\Sigma_{f(x)}$, and $\Lambda_{f;x;x'}$ via a combination of emulation and expert judgment which we now discuss in the sections that follow. In Section 5, we develop SOEF emulators for the multi-deterministic computer models; in Section 6, we make estimates for our emulators, given the collection of evaluations of the model, and construct SOEF representations for these estimates; in Section 7, we review and develop Bayes linear methods for updating means and variances for SOEF sequences, apply these methods to the sequence of emulator estimates, and deduce the appropriate revisions for the second order specification for $f^{(L)}(x)$; and in Section 8, we discuss how to make prior specifications for each of the quantities required for this analysis.

## 5. Emulation

An emulator $f(x)$ is a stochastic representation of a computer model (e.g., Bayarri et al., 2007; Kennedy et al., 2006; Goldstein and Rougier, 2006a) which is used to make probabilistic or moment based statements about model outcomes for unsampled subsets of the input space. In this paper, we have the additional purpose of emulating the latent model and learning about its second order properties as well.

A variety of approaches exist to emulate both univariate and multivariate deterministic model outcomes (e.g., Bayarri et al., 2007; Rougier, 2008; Conti and O'Hagan, 2008). We build from such approaches and emulate multi-deterministic model results with parameters that depend upon the system condition. We shall suppose that our emulator for $\{f\}_{[n]}$ is

of the form

$$f^{(c_j)}(x) \quad = \quad g(x)\beta^{(j)} + \epsilon^{(j)}(x), \tag{7}$$

$$\mathrm{E}[\epsilon^{(j)}(x)] = 0, \ \mathrm{Cov}[\epsilon^{(j)}(x), \epsilon^{(j)}(x')] = \Sigma_\epsilon K(\theta, x, x')$$

where $g(x)$ is a specified function of $x$ that results in a $1 \times r$ row vector; $\beta^{(j)}$ is a $r \times q$ matrix of model coefficients for condition $c_j$; $\epsilon^{(j)}(x)$ is a $1 \times q$ vector of error terms for condition $c_j$ that is $x$-dependent, with mean zero, uncorrelated with $\beta^{(j)}$, and modeled to have a separable covariance structure; $\Sigma_\epsilon$ is a $q \times q$ variance matrix for $\epsilon^{(j)}(x)$ that is constant over both $x$ and $c$; $K(\theta, x, x')$ represents the correlation between any two points in $\mathcal{X}$ and reflects the local smoothness of the model output $f^{(c_j)}(x)$ as a function of $\theta$; and, $\theta$ is $q \times 1$ and constant across conditions (both $\theta$ and $\Sigma_\epsilon$ are constant, because $\{f\}_{[n]}$ is SOEF). The benefit of using condition-dependent model terms $\beta^{(j)}$ and $\epsilon^{(j)}(x)$, is twofold. First, we may emulate $\{f^{(c_1)}, ..., f^{(c_m)}\}$ based on the corresponding sets of model realizations $\{f^{(c_1)}\}_{[n^{(c)}]}$, $\{f^{(c_2)}\}_{[n^{(c)}]}$, ..., $\{f^{(c_m)}\}_{[n^{(c)}]}$ separately, rather than simultaneously. This provides great computational and theoretical simplifications, provided that we have a large enough design that each model may be emulated with an acceptable degree of precision. Second, we may learn about the mean behavior of the overall multi-deterministic model and the degree to which any $c = c_j$ impacts model outcomes, including $c = L$, from the sequences $\{\beta\}_{[m]}$ and $\{\epsilon(x)\}_{[m]}$ (per $x$).

Our judgment that $\{f^{(L)}(x), f^{(c_1)}(x), , f^{(c_2)}(x), ...\}$ is SOEF (Section 4) implies that the sequences of emulator parameters, $\{\beta^{(L)}, \beta^{(1)}, \beta^{(2)}, ...\}$ and error terms $\{\epsilon^{(L)}(x), \epsilon^{(1)}(x), ...\}$ are SOE and SOEF respectively. Thus, by the SOE Representation Theorem (for the coefficients) and the SOEF Representation Theorem (for the error terms), we have decompositions

$$\begin{aligned} \beta^{(j)} \quad &= \quad \mathcal{M}\{\beta\} + \mathcal{R}_j\{\beta\} \\ \epsilon^{(j)}(x) \quad &= \quad \mathcal{M}\{\epsilon(x)\} + \mathcal{R}_j\{\epsilon(x)\} \end{aligned}$$

where $\{\mathcal{R}_1\{\beta\}, \mathcal{R}_2\{\beta\}, ...\}$ are SOE with mean zero; $\mathcal{R}_j\{\beta\}$ is uncorrelated with $\mathcal{M}\{\beta\}$; $\mathcal{R}_j\{\beta\}$ is uncorrelated with $\mathcal{R}_{j'}\{\beta\}$ for $j \neq j'$;$\{\mathcal{R}_1\{\epsilon(x)\}, \mathcal{R}_2\{\epsilon(x)\}, ...\}$ are SOEF with mean zero; $\mathcal{R}_j\{\epsilon(x)\}$ is uncorrelated with $\mathcal{M}\{\epsilon(x)\}$; and $\mathcal{R}_j\{\epsilon(x)\}$ is uncorrelated with $\mathcal{R}_{j'}\{\epsilon(x)\}$, $j \neq j'$. In turn, we have the SOEF decompositions of the emulators,

$$\mathcal{M}\{f(x)\} \quad = \quad g(x)\mathcal{M}\{\beta\} + \mathcal{M}\{\epsilon(x)\} \tag{8}$$

$$\mathcal{R}_j\{f(x)\} \quad = \quad g(x)\mathcal{R}_j\{\beta\} + \mathcal{R}_j\{\epsilon(x)\}. \tag{9}$$

Given the above decompositions, we make two points. First, emulator (7) models outcomes similar to a mixed statistical model with fixed terms $\mathcal{M}\{f(x)\}$ and random interaction terms $\mathcal{R}_j\{f(x)\}$. However, unlike mixed models, we do not presume fully exchangeable computer outcomes to assess $\mathcal{M}\{f(x)\}$ and $\mathcal{R}_j\{f(x)\}$. Second, we may extend Equations

(4)-(6), as follows:

$$\begin{aligned}
\mu_{f(x)} &= \mathrm{E}\Big[g(x)\mathcal{M}\{\beta\} + \mathcal{M}\{\epsilon(x)\}\Big] = g(x)\mu_\beta + \mu_{\epsilon(x)} && (10)\\
\Sigma_{\mathcal{M}\{f(x)\}} &= \mathrm{Var}\Big[g(x)\mathcal{M}\{\beta\} + \mathcal{M}\{\epsilon(x)\}\Big] \\
&= g(x)\Sigma_{\mathcal{M}\{\beta\}}g(x)^T + \Sigma_{\mathcal{M}\{\epsilon(x)\}} && (11)\\
\Sigma_{\mathcal{R}\{f(x)\}} &= \mathrm{Var}\Big[g(x)\mathcal{R}_j\{\beta\} + \mathcal{R}_j\{\epsilon(x)\}\Big] \\
&= g(x)\Sigma_{\mathcal{R}\{\beta\}}g(x)^T + \Sigma_{\mathcal{R}\{\epsilon(x)\}} && (12)\\
\Lambda_{f;x;x'} &= \mathrm{Cov}\Big[g(x)\mathcal{M}\{\beta\} + \mathcal{M}\{\epsilon(x)\}, g(x')\mathcal{M}\{\beta\} + \mathcal{M}\{\epsilon(x')\}\Big] \\
&= g(x)\Sigma_{\mathcal{M}\{\beta\}}g(x')^T + \Lambda_{\mathcal{M}\{\epsilon;x;x'\}}. && (13)
\end{aligned}$$

We discuss in the next section methods for both fitting a computer model emulator and including estimation error in assessments of $\mu_{f(x)}$, $\Sigma_{\mathcal{M}\{f(x)\}}$, $\Sigma_{\mathcal{R}\{f(x)\}}$, and $\Lambda_{f;x;x'}$.

## 6. Emulator Estimation

In order to fit the emulator (7), given the computer model evaluations, we use generalized least squares (GLS) methods (Appendix A). GLS has two desirable effects that greatly simplify all of the following calculations in this section. First, it de-correlates the error terms within a system condition. Secondly, the resulting estimates for the coefficients and residuals are themselves SOE and SOEF because we are using a balanced design for the computer evaluations.

Let $\hat{f}^{(c_j)}(x)$ represent the estimated version of the computer model emulator for system condition $c_j$, and let $\{\widehat{\beta}\}_{[m]}$ and $\{\widehat{\epsilon}\}_{[n]}$ represent the model terms in $\{\hat{f}(x)\}_{[m]}$ that include estimation error $e(\cdot)$,

$$\widehat{\beta}^{(j)} = \beta^{(j)} + e(\widehat{\beta}^{(j)}), \qquad \widehat{\epsilon}^{(j)}(x) = \epsilon^{(j)}(x) + e(\widehat{\epsilon}^{(j)}(x)),$$

where, for $a \in (\beta, \epsilon(x))$, estimation error $e(\widehat{a}^{(j)})$ equals $\widehat{a}^{(j)} - a^{(j)}$. Since we are using GLS estimation for a balanced design across the conditions, $\widehat{\beta}$ and $\widehat{\epsilon}(x)$ have the following properties: for $a \in (\beta, \epsilon(x))$, $\mathrm{E}[e(\widehat{a}^{(j)})] = 0$, $\mathrm{Var}[e(\widehat{a}^{(j)})] = \Sigma_{e(\widehat{a})}$ is constant, $\mathrm{Cov}[a^{(j)}, e(\widehat{a}^{(j)})] = 0$, $\mathrm{Cov}[e(\widehat{a}^{(j)}), e(\widehat{a}^{(j')})] = 0$ for $j \neq j'$, and the sequences of estimates $\{\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}, ...\}$, and $\{\{\widehat{\epsilon}^{(1)}\}_{[n^{(c)}]}, \{\widehat{\epsilon}^{(2)}\}_{[n^{(c)}]}, ...\}$ are SOE and SOEF respectively. In turn,

$$\begin{aligned}
\widehat{\beta}^{(j)} &= \mathcal{M}\{\widehat{\beta}\} + \mathcal{R}_j\{\widehat{\beta}\} &&= \mathcal{M}\{\beta\} + \mathcal{R}_j\{\widehat{\beta}\} \\
&&&= \mathcal{M}\{\beta\} + \big(\mathcal{R}_j\{\beta\} + e(\widehat{\beta}^{(j)})\big) \\
\widehat{\epsilon}^{(j)}(x) &= \mathcal{M}\{\widehat{\epsilon}(x)\} + \mathcal{R}_j\{\widehat{\epsilon}(x)\} &&= \mathcal{M}\{\epsilon(x)\} + \mathcal{R}_j\{\widehat{\epsilon}(x)\}, \\
&&&= \mathcal{M}\{\epsilon(x)\} + \big(\mathcal{R}_j\{\epsilon(x)\} + e(\widehat{\epsilon}^{(j)}(x))\big).
\end{aligned}$$

and $\{\hat{f}^{(c_1)}(x), \hat{f}^{(c_2)}(x), ...\}$ is SOEF. Also, the means of $f^{(c_j)}(x)$ and $\hat{f}^{(c_j)}(x)$ are equal because $\mathcal{M}\{\beta\} = \mathcal{M}\{\widehat{\beta}\}$ and $\mathcal{M}\{\epsilon(x)\} = \mathcal{M}\{\widehat{\epsilon}(x)\}$,

$$\mu_{\hat{f}(x)} = \mu_{f(x)} \qquad \Sigma_{\mathcal{M}\{\hat{f}(x)\}} = \Sigma_{\mathcal{M}\{f(x)\}};$$

but, the variance of $f^{(j)}(x)$ is less than the variance of $\hat{f}^{(c_j)}(x)$ because

$$\Sigma_{\widehat{\mathcal{R}}\{f(x)\}} = \Sigma_{\mathcal{R}\{f(x)\}} + g(x)\Sigma_{e(\widehat{\beta})}g(x)^T + \Sigma_{e(\widehat{\epsilon}(x))}.$$

To account for the change in variance, we describe in the next section the Bayes linear adjustments of the prior expectation and variance of $f^{(L)}(x)$ given $\{\widehat{\beta}\}_{[m]}$ and $\{\widehat{\epsilon}\}_{[n]}$, rather $\{\beta\}_{[m]}$ and $\{\epsilon\}_{[n]}$.

## 7. Bayes Linear Adjustments for the Latent Emulator

Since we are only willing to assume that the models are second-order exchangeable, we use Bayes linear methodology that relies on expectation as the primitive, not probability. Thus, without distributional specifications, Bayes linear methods entail adjusting prior moments for quantities of interest by data. In this paper, we aim to adjust the moments of the latent model $f^{(L)}(x)$ given a collection of realized models and their emulator estimates.

Using standard, Bayes linear notation, the *adjusted* expectation and *adjusted* variance are denoted as follows: given random vectors $a$ and $b$, let $\mathrm{E}_b[a]$ and $\mathrm{Var}_b[a]$ represent the adjusted expectation and variance of $a$ given $b$. These quantities are derived from first eliciting the joint prior expectation, variance and covariance for $(a, b)$, and subsequently applying the Bayes linear equations,

$$\mathrm{E}_b[a] = \mathrm{E}[a] + \mathrm{Cov}[a,b]\mathrm{Var}[b]^{-1}\Big(b - \mathrm{E}[b]\Big) \quad (14)$$

$$\mathrm{Var}_b[a] = \mathrm{Var}[a] - \mathrm{Cov}[a,b]\mathrm{Var}[b]^{-1}\mathrm{Cov}[b,a]; \quad (15)$$

a detailed discussion of the Bayes linear approach is given in (Goldstein and Woof, 2007). The specification of the joint prior moments is an important step in Bayes linear analyses. Thus, in the sections that follow we include detailed descriptions about how to make the needed specifications within the context of multi-deterministic computer experiments.

### 7.1. Expectation

Goldstein and Woof (2007, pg. 196) show that a sample mean for a SOE sequence is Bayes linear sufficient for both: (i) the population mean (i.e. Bayes linear adjustments of the population mean given either the full sample or just the sample mean are equal), and (ii) any further members of the SOE sequence which are not included in the sample. Since the sequences $\{\widehat{\beta}\}_{[m]}$ and $\{\{\widehat{\epsilon}^{(1)}\}_{[n^{(c)}]}, ..., \{\widehat{\epsilon}^{(m)}\}_{[n^{(c)}]}\}$ are SOE and SOEF respectively, we know that $\widehat{\overline{\beta}}$ and $\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}$,

$$\widehat{\overline{\beta}} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\beta}^{(j)}, \qquad \widehat{\overline{\epsilon}}(x) = \frac{1}{m}\sum_{j=1}^{m}\widehat{\epsilon}^{(j)}(x),$$

are Bayes linear sufficient to update $\mathcal{M}\{\beta\}$ and $\mathcal{M}\{\epsilon(x)\}$. The prior moments of $\widehat{\overline{\beta}}$ and $\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}$ are

$$\begin{array}{ll} \mathrm{E}[\widehat{\overline{\beta}}] = \mu_\beta & \mathrm{Var}[\widehat{\overline{\beta}}] = \Sigma_{\mathcal{M}\{\beta\}} + \frac{1}{m}(\Sigma_{\mathcal{R}\{\beta\}} + \Sigma_{e(\widehat{\beta})}) \\ \mathrm{E}[\widehat{\overline{\epsilon}}(x)] = \mu_{\epsilon(x)} & \mathrm{Var}[\widehat{\overline{\epsilon}}(x)] = \Sigma_{\mathcal{M}\{\epsilon(x)\}} + \frac{1}{m}(\Sigma_{\mathcal{R}\{\epsilon(x)\}} + \Sigma_{e(\widehat{\epsilon}(x))}), \end{array}$$

and, the adjusted moments of $f^{(L)}(x)$ given $\widehat{\overline{\beta}}$ and $\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}$ are

$$\mathrm{E}_{(\{\widehat{\beta}\}_{[m]},\{\widehat{\epsilon}\}_{[n]})}[f^{(L)}(x)] \equiv \mathrm{E}_{\widehat{\overline{\beta}},\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}\}}[f^{(L)}(x)]$$

$$\begin{aligned} &= g(x)\mathrm{E}_{\{\widehat{\overline{\beta}},\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}\}}[\mathcal{M}\{\beta\}] + \mathrm{E}_{\{\widehat{\overline{\beta}},\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}\}}[\mathcal{M}\{\epsilon(x)\}], \\ &\approx g(x)\mathrm{E}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}] + \mathrm{E}_{\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}] \end{aligned} \tag{16}$$

$$\begin{aligned} \mathrm{Var}_{(\{\widehat{\beta}\}_{[m]},\{\widehat{\epsilon}\}_{[n]})}[f^{(L)}(x)] &\equiv \mathrm{Var}_{\{\widehat{\overline{\beta}},\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}\}}[f^{(L)}(x)] \\ &\approx g(x)\mathrm{Var}_{\widehat{\overline{\beta}}}\Big[\mathcal{M}\{\beta\} + \mathcal{R}_L\{\beta\}\Big]g(x)^T + \mathrm{Var}_{\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}}\Big[\mathcal{M}\{\epsilon(x)\} + \mathcal{R}_L\{\epsilon(x)\}\Big] \\ &= g(x)\mathrm{Var}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}]g(x)^T + g(x)\Sigma_{\mathcal{R}\{\beta\}}g(x)^T \\ &\quad + \mathrm{Var}_{\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}] + \Sigma_{\mathcal{R}\{\epsilon(x)\}}, \end{aligned} \tag{17}$$

where $\mathrm{E}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}]$, $\mathrm{E}_{\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}]$, $\mathrm{Var}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}]$, and $\mathrm{Var}_{\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}]$ are the adjusted expectations and variances of $\mathcal{M}\{\beta\}$ and $\mathcal{M}\{\epsilon(x)\}$ as defined by Equations (14) and (15). Equations (16) and (17) are approximations because we zero the natural correlation between the $\widehat{\overline{\beta}}$ and $\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}$ that is induced by their estimation errors. Goldstein and Rougier (2006b) make the same approximation, and for large samples, this will have little impact on model based inferences.

Notice that two terms in Equation (17) are independent of both $\widehat{\overline{\beta}}$ and $\{\widehat{\overline{\epsilon}}\}_{[n^{(c)}]}$; the latent residuals $\mathcal{R}_L\{\beta\}$ and $\mathcal{R}_L\{\epsilon(x)\}$ are, by the SOE and SOEF representation theorems respectively, orthogonal to the sample means. Thus, Bayes linear adjustments by the sample means do not affect residual variances $\Sigma_{\mathcal{R}\{\beta\}}$ and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$. In the next section, we show how to make data adjusted assessments of $\Sigma_{\mathcal{R}\{\beta\}}$ and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$.

### 7.2. Residual Variances

Let $S_{\widehat{\beta}}$ and $S_{\widehat{\epsilon}(x)}$ represent the observed sample variances of $\{\widehat{\beta}\}_{[m]}$ and $\{\widehat{\epsilon}(x)\}_{[m]}$ respectively. In this section, we use $S_{\widehat{\beta}}$ and $S_{\widehat{\epsilon}(x)}$ to carry out a Bayes linear adjustment of $\Sigma_{\mathcal{R}\{\beta\}}$ and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$. To do so however, we make two additional second-order exchangeable assumptions and decompose $S_{\widehat{\beta}}$ and $S_{\widehat{\epsilon}(x)}$ so that the joint prior moment assessments of $(S_{\widehat{\beta}}, \Sigma_{\mathcal{R}\{\beta\}})$ and $(S_{\widehat{\epsilon}(x)}, \Sigma_{\mathcal{R}\{\epsilon(x)\}})$ are clear.

#### 7.2.1. SOE and SOEF Assumptions for the Squared Residuals

The additional assumptions pertain to the $m+1$ sequences of squared residuals of the emulator terms which, for ease in notation, we denote as $\{v_\beta\}_{[m,L]}$ and $\{v_{\epsilon(x)}\}_{[m,L]}$ that contain elements $v_\beta^{(j)}$ and $v_{\epsilon(x)}^{(j)}$,

$$v_\beta^{(j)} = \mathcal{R}_j\{\beta\}\mathcal{R}_j\{\beta\}^T \quad \text{and} \quad v_{\epsilon(x)}^{(j)} = \mathcal{R}_j\{\epsilon(x)\}\mathcal{R}_j\{\epsilon(x)\}^T.$$

If we assume that the $\{v_\beta^{(L)}, v_\beta^{(1)}, v_\beta^{(2)}, ...\}$ is SOE and $\{v_{\epsilon(x)}^{(L)}, v_{\epsilon(x)}^{(1)}, v_{\epsilon(x)}^{(2)}, ...\}$ is SOEF, then we have decompositions,

$$v_\beta^{(j)} = \mathcal{M}\{v_\beta\} + \mathcal{R}_j\{v_\beta\} \quad \text{and} \quad v_{\epsilon(x)}^{(j)} = \mathcal{M}\{v_{\epsilon(x)}\} + \mathcal{R}_j\{v_{\epsilon(x)}\}$$

that satisfy properties a-c of the SOEF representation theorem. In particular,

$$\begin{aligned} \mathrm{E}[v_\beta^{(j)}] &= \mu_{v_\beta} \qquad \mathrm{Var}[v_\beta^{(j)}] = \Sigma_{\mathcal{M}\{v_\beta\}} + \Sigma_{\mathcal{R}\{v_\beta\}} \\ \mathrm{E}[v_{\epsilon(x)}^{(j)}] &= \mu_{v_{\epsilon(x)}} \qquad \mathrm{Var}[v_{\epsilon(x)}^{(j)}] = \Sigma_{\mathcal{M}\{v_{\epsilon(x)}\}} + \Sigma_{\mathcal{R}\{v_{\epsilon(x)}\}}. \end{aligned}$$

Since $\mathrm{E}[\mathcal{R}_j\{\beta\}] = 0$ and $\mathrm{E}[\mathcal{R}_j\{\epsilon(x)\}] = 0$ for any $j \in \{L, 1, 2...\}$, the variances of $\mathcal{R}_j\{\beta\}$ and $\mathcal{R}_j\{\epsilon(x)\}$ equal the means of $v_\beta$ and $v_{\epsilon(x)}$ respectively,

$$\begin{aligned} \mathrm{Var}[\mathcal{R}_j\{\beta\}] &= \Sigma_{\mathcal{R}\{\beta\}} = \mu_{v_\beta} \\ \mathrm{Var}[\mathcal{R}_j\{\epsilon(x)\}] &= \Sigma_{\mathcal{R}\{\epsilon(x)\}} = \mu_{v_{\epsilon(x)}}. \end{aligned}$$

Thus, updating the residual variances $\Sigma_{\mathcal{R}\{\beta\}}$ and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$ is similar to adjusting the expectations of $v_\beta^{(j)}$ and $v_{\epsilon(x)}^{(j)}$ by $S_{\widehat{\beta}}$ and $S_{\widehat{\epsilon}(x)}$ respectively for any $j$; i.e., we assess $\mathrm{E}_{S_{\widehat{\beta}}}[\mathcal{M}\{v_\beta\}]$ and $\mathrm{E}_{S_{\widehat{\epsilon}(x)}}[\mathcal{M}\{v_{\epsilon(x)}\}]$).

### *7.2.2. Marginal and Joint Prior Variance Specifications*

Adjustments $\mathrm{E}_{S_{\widehat{\beta}}}[\mathcal{M}\{v_\beta\}]$ and $\mathrm{E}_{S_{\widehat{\epsilon}(x)}}[\mathcal{M}\{v_{\epsilon(x)}\}]$ rely on the marginal and joint prior expectations and variances of $(v_\beta^{(j)}, S_{\widehat{\beta}})$ and $(v_{\epsilon(x)}^{(j)}, S_{\widehat{\epsilon}})$. However, eliciting the partial priors for $v_\beta^{(j)}$ and $v_{\epsilon(x)}^{(j)}$ from experts can be challenging. Experts are constrained by the fact that the final adjusted variances must be positive semi-definite and, given multivariate outputs, the emulator coefficients and error terms have multi-dimensional variance arrays that can be hard to conceptualize. Therefore, we suggest to either vectorize the variance arrays or model the multivariate output univariately. In doing so, we need only consider standard, two-dimensional variance matrices for our variance adjustment of $f^{(L)}(x)$ which can be assessed by a *semi-adjustment* technique developed in Goldstein and Woof (2007, pg. 286).

In short, the Bayes linear semi-adjustment technique separates the tasks of learning from the data the diagonal and off-diagonal elements within a variance matrix. To assess the column vector of univariate variances (the diagonal terms of the variance matrix), standard Bayes linear updating applies; to estimate the posterior correlation matrix, a weighted average of a prior and estimated correlation matrix is taken. In turn, the updated correlation matrix and the outer product of the square root vector of updated, univariate variances are multiplied (element-wise). The result is a data semi-adjusted variance matrix. In Section 9, we use the semi-adjustment method to assess the variance matrix of $\beta$ and rely on data collected from a pilot study to specify a prior correlation matrix.

Because we rely on GLS to emulate the computer model, we do not need to use the semi-adjustment technique to assess $\mathrm{Var}[\epsilon^{(j)}(x_i)]$. Due to the data rotation and modeling assumptions discussed in Section 6, $\mathrm{Corr}[\epsilon^{(j)}(x_i), \epsilon^{(j')}(x_{i'})] = 0$ and $\mathrm{Var}[\epsilon^{(j)}(x_i)] = \mathrm{Var}[\epsilon^{(j')}(x_{i'})] = \Sigma_\epsilon$ for any $j$, $j'$, $i$, and $i'$, and $\Sigma_\epsilon$ has only non-zero elements on the diagonal. Thus, we need only adjust one, univariate variance to characterize $\Sigma_\epsilon$ which requires the joint prior specifications of $(v_{\epsilon(x)}^{(j)}, S_{\widehat{\epsilon}})$. The remainder of this section describes an approach for specifying the joint prior expectation and variance for both $(v_{\epsilon(x)}^{(j)}, S_{\widehat{\epsilon}})$ and $(v_\beta^{(j)}, S_{\widehat{\beta}})$.

The joint specifications follow clearly from the decompositions $S_{\widehat{\beta}}$ and $S_{\widehat{\epsilon}(x)}$ as defined in Goldstein and Woof (2007, pg. 282) and applied to exchangeable sequences with estimation error:

$$S_{\widehat{\beta}} = \mathcal{M}\{v_\beta\} + T_{\widehat{\beta}} \quad \text{and} \quad S_{\widehat{\epsilon}(x)} = \mathcal{M}\{v_{\epsilon(x)}\} + T_{\widehat{\epsilon}(x)},$$

where for $a \in [\beta, \epsilon(x)]$,

$$T_{\hat{a}} \;=\; \Sigma_{e(\hat{a})} + \frac{1}{m}\sum_j \mathcal{R}_j\{v_a\} - \frac{1}{m}\sum_{j\neq k} \mathcal{R}_j\{a\}\mathcal{R}_k\{a\}^T,$$

and $T_{\widehat{\beta}}$ and $T_{\widehat{\epsilon}(x)}$ have the following properties (Goldstein and Woof, 2007, pg. 282, eq. 8.7): $\mathrm{E}[T_{\widehat{\beta}}] = \Sigma_{e(\widehat{\beta})}$, $\mathrm{E}[T_{\widehat{\epsilon}(x)}] = \Sigma_{e(\widehat{\epsilon}(x))}$, $\mathrm{Var}[T_{\widehat{\beta}}] = \Sigma^\dagger_{e(\widehat{\beta})} + \Sigma_{T_\beta}$, $\mathrm{Var}[T_{\widehat{\epsilon}(x)}] = \Sigma^\dagger_{e(\widehat{\epsilon}(x))} + \Sigma_{T_{\epsilon(x)}}$, $\mathrm{Cov}[\mathcal{M}\{v_\beta\}, T_{\widehat{\beta}}] = 0$, and $\mathrm{Cov}[\mathcal{M}\{v_{\epsilon(x)}\}, T_{\widehat{\epsilon}(x)}] = 0$, where $\Sigma^\dagger_{e(\widehat{a})}$ is the variance of $e(\widehat{a})^2$ which can be thought of as the variance of the variance of $e(\widehat{a})$. As a result, the joint prior expectations and variances of $(v^{(j)}_\beta, S_{\widehat{\beta}})$ and $(v^{(j)}_{\epsilon(x)}, S_{\widehat{\epsilon}})$ are

$$\begin{aligned}
&\mathrm{E}[v^{(j)}_\beta, S_\beta] = [\mu_{v_\beta}, \mu_{v_\beta} + \Sigma_{e(\widehat{\beta})}] && \mathrm{E}[v^{(j)}_{\epsilon(x)}, S_{\epsilon(x)}] = [\mu_{v_{\epsilon(x)}}, \mu_{v_{\epsilon(x)}} + \Sigma_{e(\widehat{\epsilon}(x))}]\\
&\mathrm{Var}[S_{\widehat{\beta}}] = \Sigma_{\mathcal{M}\{v_\beta\}} + \Sigma^\dagger_{e(\widehat{\beta})} + \Sigma_{T_\beta} && \mathrm{Var}[S_{\widehat{\epsilon}(x)}] = \Sigma_{\mathcal{M}\{v_\epsilon\}} + \Sigma^\dagger_{e(\widehat{\epsilon}(x))} + \Sigma_{T_{\epsilon(x)}}\\
&\mathrm{Var}[v^{(j)}_\beta] = \Sigma_{\mathcal{M}\{v_\beta\}} + \Sigma_{\mathcal{R}\{v_\beta\}} && \mathrm{Var}[v^{(j)}_{\epsilon(x)}] = \Sigma_{\mathcal{M}\{v_{\epsilon(x)}\}} + \Sigma_{\mathcal{R}\{v_{\epsilon(x)}\}}\\
&\mathrm{Cov}[v^{(j)}_\beta, S_{\widehat{\beta}}] = \Sigma_{\mathcal{M}\{v_\beta\}} && \mathrm{Cov}[v^{(j)}_{\epsilon(x)}, S_{\widehat{\epsilon}}] = \Sigma_{\mathcal{M}\{v_\epsilon\}},
\end{aligned}$$

and adjustments $\mathrm{E}_{S_{\widehat{\beta}}}[\mathcal{M}\{v_\beta\}]$ and $\mathrm{E}_{S_{\widehat{\epsilon}(x)}}[\mathcal{M}\{v_{\epsilon(x)}\}]$ may follow by applying formula (14). These adjustments, in addition to those described in Section 7.1, enable the calculation of a data-informed variance.

### 7.3. Bayes Linear Two-stage Analysis

Our analysis of $f^{(L)}(x)$ proceeds in two stages (Goldstein and Woof, 2007, pg. 288). For stage one, we adjust the residual variances by the data via $\mathrm{E}_{S_{\widehat{\beta}}}[\mathcal{M}\{v_\beta\}]$ and $\mathrm{E}_{S_{\widehat{\epsilon}(x)}}[\mathcal{M}\{v_{\epsilon(x)}\}]$; and, for stage two, we use these adjustments as plug-in estimates for $\Sigma_{\mathcal{R}\{\beta\}}$ and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$ to assess $\mathrm{Var}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}]$, $\mathrm{Var}_{\{\widehat{\epsilon}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}]$, and Equation (16). We denote the one and two-stage adjustments by superscripts '(1)' and '(2)' respectively and define the two-stage adjustments of $\mathrm{E}[f^{(L)}(x)]$ and $\mathrm{Var}[f^{(L)}(x)]$ as follows:

$$\mathrm{E}^{(2)}_D[f^{(L)}(x)] \;\approx\; g(x)\mathrm{E}^{(2)}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}] + \mathrm{E}^{(2)}_{\{\widehat{\epsilon}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}] \tag{18}$$

$$\begin{aligned}
\mathrm{Var}^{(2)}_D[f^{(L)}(x)] \;\approx\;& g(x)\mathrm{Var}^{(2)}_{\widehat{\overline{\beta}}}[\mathcal{M}\{\beta\}]g(x)^T + g(x)\mathrm{E}^{(1)}_{S_{\widehat{\beta}}}[\mathcal{M}\{v_\beta\}]g(x)^T\\
&+\mathrm{Var}^{(2)}_{\{\widehat{\epsilon}\}_{[n^{(c)}]}}[\mathcal{M}\{\epsilon(x)\}] + \mathrm{E}^{(1)}_{S_{\widehat{\epsilon}(x)}}[\mathcal{M}\{v_{\epsilon(x)}\}],
\end{aligned} \tag{19}$$

where $D = \{\widehat{\overline{\beta}}, \{\widehat{\epsilon}\}_{[m]}, S_{\widehat{\beta}}, S_{\widehat{\epsilon}}\}$.

## 8. Prior Specifications

Some of the prior quantities needed for adjustments (18) and (19) might be somewhat unfamiliar to experts and, as discussed by O'Hagan (1998), reasonable prior specifications are vital for Bayesian analyses. In this section, we describe some examples of the types of modeling decisions which we make to simplify the specifications.

For a random SOE quantity $a$, we might be prepared to specify directly one of the following three variances: $\Sigma_a$, $\Sigma_{\mathcal{M}\{a\}}$, or $\Sigma_{\mathcal{R}\{a\}}$, where $\Sigma_a = \Sigma_{\mathcal{M}\{a\}} + \Sigma_{\mathcal{R}\{a\}}$. To deduce the remaining variance terms, we presume that $\Sigma_{\mathcal{M}\{a\}} = w_0 + w_1\Sigma_{\mathcal{R}\{a\}}$ where $w_0$ and $w_1$ are expert elicited, and in Goldstein, House, and Rougier (2008) we discuss a complex method for making such judgments. If $w_0$ is judged to equal zero, then the procedure for specifying $w_1$ simplifies. The variance $\Sigma_{\mathcal{M}\{a\}}$ is strictly proportional to $\Sigma_{\mathcal{R}\{a\}}$, and the value for $w_1$ should be small when the expert believes that the realized models outcomes are indicative of the entire set of possible model outcomes (and large otherwise). In Section 9, we make the judgment that $w_0 = 0$ and elicit values for $w_1 < 1$ from experts which we denote by $w_{1a}$ and $w_{1b}$.

We make similar proportionality assumptions as suggested in Goldstein and Woof (2007) to solve for the variance of the mean variance and the variance of the residual variance, e.g., for random quantity $a$, $\Sigma_{\mathcal{M}\{v_a\}} = w_2\mu_{v_a}^2$ and $\Sigma_{\mathcal{R}\{v_a\}} = w_3(\Sigma_{\mathcal{M}\{v_a\}} + \mu_{v_a}^2)$. The constant $w_2$ can be found in terms of $m$ and two expert elicited quantities $p$ and $k$,

$$w_2 = \frac{k(1-p)}{(k(p-1) + p(m-1))}.$$

The first quantity $p$ pertains to the proportion of the updated variance of $a$ we expect prior judgments to explain. The second quantity $k$ is a function of the expected kurtosis for the distribution of $a$ ($\mathrm{Kur}(a)$) where

$$k = \frac{1}{m}\Big[(m-1)\big(\mathrm{Kur}(a) - 1\big) + 2\Big].$$

Given $\mathrm{Kur}(a)$, we may solve for $w_3$,

$$w_3 = \mathrm{Kur}(a) - 1.$$

As mentioned previously, the specifications of prior moments is an important component in Bayes linear analyses. In the next section, we apply our methods and exemplify how we elicit prior information from both experts and data collected previously.

## 9. Application: Galform

The Galform model was developed by the Durham Semi-analytical modeling group at the Institute of Computational Cosmology (ICC), the University of Durham, United Kingdom and simulates the creation and evolution of approximately one million galaxies, from the beginning of the universe until the current day. It relies on complex mathematical equations which characterize relevant physical processes, including gravitational collapse, radioactive cooling and star formation, to predict (among other variables) the number and luminosity of the simulated galaxies that are observed in our universe. The Galform equations require

**Table 1.** List of input variable ($x = [x_1 ... x_8]$) abbreviations with corresponding minimum and maximum values designed for Galform.

| $l : x_l$ | Variable | Minimum | Maximum |
|---|---|---|---|
| 1 | *vhotdisk* | 100 | 550 |
| 2 | *alpha_reheat* | 0.20 | 1.20 |
| 3 | *alpha_cool* | 0.20 | 1.20 |
| 4 | *vhotburst* | 100 | 550 |
| 5 | *epsilon_Star* | 10 | 1000.00 |
| 6 | *stabledisk* | 0.65 | 0.95 |
| 7 | *alphahot* | 2.00 | 3.70 |
| 8 | *yield* | 0.02 | 0.05 |

the specification of values for $p = 8$ input parameters† which are listed in abbreviated form in Table 1. Further details concerning the input parameters and the model itself can be found in several references including Cole et al. (1994); Springel et al. (2005); Bower et al. (2006); Baugh (2006), and references therein.

One output $f(x)$ produced by Galform is a $q$-dimensional vector ($f(x) = \{f_1(x), ..., f_q(x)\}$) which effectively represents the proportion (logarithm base 10) of all galaxies simulated per unit volume $\phi(\mathrm{h}^3\ \mathrm{Mpc}^{-3}\mathrm{mag}^{-1})$ that fall within $q = 35$, mutually exclusive ranges of $b_j$-band luminosity‡ or absolute magnitude ($-\mathrm{M}_{bj} + 5\log_{10}\mathrm{h}$). Figure 1 includes one example of an output from Galform. Six vertical lines are drawn in Figure 1 to mark six luminosities or points of interest on the curves that are particularly sensitive to the Galform inputs. These points include 17, 18.25, 20, 21, 21.75, and 22.25 ($-\mathrm{M}_{bj} + 5\log_{10}\mathrm{h}$).

An important component to the formation of our universe is the presence of dark matter (Figure 2). Thus, coded into Galform is a system condition variable $c$ called *region* for dark matter specifications. The name *region* is a natural choice because another computer model simulated one large configuration of dark matter (known as, the "millennium simulation") that was subdivided into 512 spatial regions, each of which can be used as forcing functions for Galform.

### 9.1. Experimental Design and Data Transformations

The first $m = 40$ of the 512 regions were preselected and fixed for a Galform computer experiment. An 8-dimensional Latin hypercube experimental design was used to select $n^{(c)} = 1000$ input values $x_i$, $i \in 1, ..., n^{(c)}$. Since the same $n^{(c)}$ values were input into Galform for each pre-selected region or system condition $j$ where $j \in [1, ..., m]$, the Galform computer experiment included $n = 40,000$ design points.

For the following analysis, incomplete runs were removed which reduced $n^{(c)}$ to 967 (runs only failed at random due to computational, not theoretical, limitations), and, per

†The current version of Galform actually includes 17 input parameters. However, 8 of the 17 variables influence the luminosity predictions substantially more than the remaining 9. Thus, the computer experiment designed and discussed in this paper set 9 input parameters to expert elicited values and varies the influential 8 parameters.

‡Galform actually makes predictions for 109 mutually exclusive luminosity ranges, but typical analysis focus on galaxies with luminosity is less than -15.5. Additionally, the $\mathrm{b}_j$-band luminosity curves is one of several outputs produced by Galform.

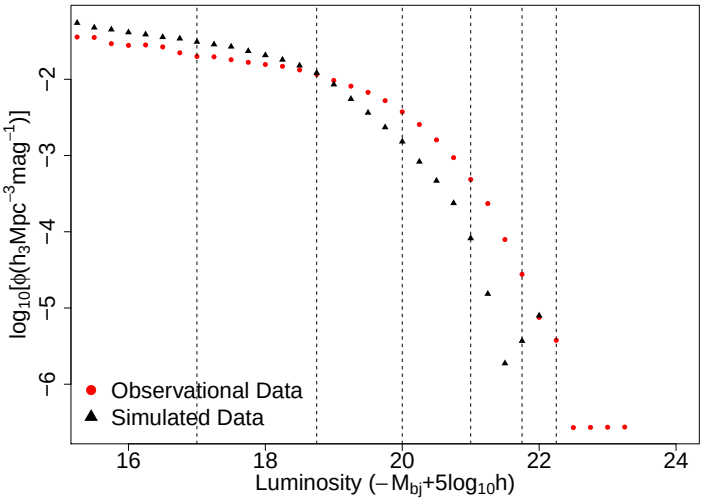



**Fig. 1.** Example of one Galform result compared to observational data. The x-axis represents the absolute magnitude of the galaxies which relates directly to the intrinsic luminosity of the galaxy, and the y-axis is the logarithm base 10 proportion of galaxies (for a given curve) per unit volume that fall within the given luminosity range.

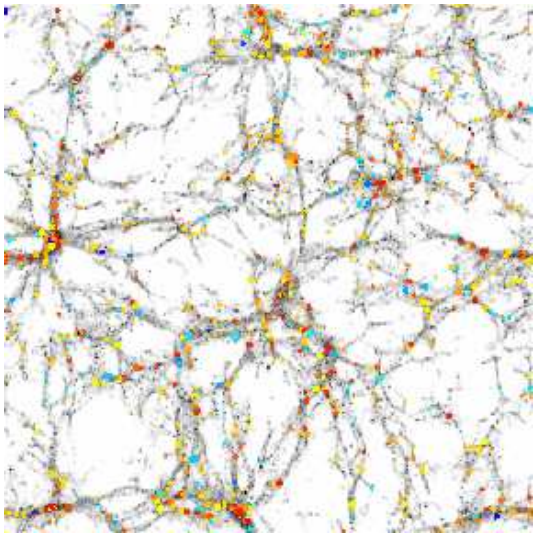


**Fig. 2.** The gray net of matter represents an example of one dark matter specification, and the dots represent different galaxies simulated with various luminosities.

luminosity, the outcomes were scaled by the mean and standard deviation across runs and regions. Additionally, each input was transformed so that $x \in [-1, 1]$.

### *9.2. Assess $f^{(L)}(x)$*

For exemplary purposes, we start by selecting one point from the luminosity curve and applying our methods as explained in Sections 4-7. Thus, our initial goal for this section is to assess the latent Galform outcome for luminosity 18.25 (-$M_{bj} + 5\log_{10}$h) which we denote by $f^{(L)}_{18.25}(x)$, and we implement our analysis of $f^{(L)}_{18.25}(x)$ in the following order:

- 9.2.1 Judge whether Galform outcomes are SOEF.
- 9.2.2 Fit a mean-emulator to learn $g(x)$ and $K$ for emulator (7).
- 9.2.3 Fit emulator (7) and store both the sequence of model coefficients $\{\beta\}_{[m]}$ and sequence of error terms $\{\epsilon\}_{[n]}$.
- 9.2.4 Obtain the data summaries and elicit the prior quantities needed to calculate the adjusted variances of $\beta^{(L)}$ and $\{\epsilon^{(L)}\}_{[n^{(c)}]}$ and the adjusted expectations of $\beta^{(L)}$ and $\{\epsilon^{(L)}\}_{[n^{(c)}]}$.
- 9.2.5 Assess results.

In Section 9.2.6, we repeat steps 9.2.1-9.2.5 for additional points on the luminosity curve; i.e., we derive the adjusted expectations and variances of $f^{(L)}_{17}(x)$, $f^{(L)}_{20}(x)$, $f^{(L)}_{21}(x)$, $f^{(L)}_{21.75}(x)$, and $f^{(L)}_{22.25}(x)$. In conclusion, we assess $f^{(L)}(x)$ based on the six sets of adjusted moments.

#### *9.2.1. SOEF Judgment*

Each specification for *region* is the result of one dark matter simulation and an underlying modeling structure may exist which correlates certain dark matter predictions more strongly than others. However, we still make the prior judgment that the $m+1$ sequence $\{f(x)\}_{[m]} = (f^{(L)}(x), \{f(x)\}_{[m]})$ is SOEF because the possible structure is, for all practical purposes, unquantifiable and unknown. The task of cataloging and parameterizing the differences in dark matter specification such as Figure 2 is enormously complicated, and even the spatial orientation of the selected regions is undocumented. In light of these limitations, a simple thought experiment conducted by an expert resulted in the judgment that the mean and correlation of Galform output given two different, randomly chosen regions is constant. Hence, we consider the sequence $\{f\}_{[n]}$ to be SOEF. Similarly, we make the provisional assumption that the simulated estimates of Galform are SOEF with the results conditional on the true dark matter; a priori, $(\{f^{(L)}\}_{[n^{(c)}]}, \{f\}_{[n]})$ are also SOEF.

#### *9.2.2. Learn Correlation Distance and $g(x)$ from Mean-Emulator*

Emulator (7) relies on specifications of $g(x)$ and $\theta$ which we choose to learn from a mean-emulator $\overline{f}(x)$,

$$\overline{f}(x) \quad = \quad g(x)\overline{\beta} + \overline{\epsilon}(x), \quad \text{Cov}[\overline{\epsilon}(x), \overline{\epsilon}(x')] = \Sigma_{\overline{\epsilon}} K(\theta, x, x'). \tag{20}$$

where, $g(x)$ and $K(\theta, x, x')$ have equivalent definitions to Equation (7); $\overline{\beta}$ is a $r \times q$ matrix of model coefficients for the mean emulator; $\overline{\epsilon}(x)$ is a Gaussian process error term; and $\Sigma_{\overline{\epsilon}}$ is a $q \times q$ matrix that represents the variance for $\overline{\epsilon}(x)$ per input $x$ and does not depend on $x$.

The advantage of using a mean-emulator (20) is that the procedures implemented to learn $g(x)$ and $K()$, such as backward stepwise regression or variograms respectively, need only apply to one sequence, i.e., the $n^{(c)}-$sequence of means $\{\overline{f}\}_{[n^{(c)}]}$, rather than the complete $m \times n^{(c)}$ collection of model runs. The implication of this choice is that the mean effects are parameterized similarly to the region effects. Allowing $g(x)$ to vary with region would avoid this restriction and enable a more complex parameterization for the effect of region than $\mathcal{M}\{f(x)\}$. However, preliminary analyses of Galform suggest that this complexity is not necessary for our research question.

There are a variety of options for $K()$, $g(x)$, and methods to fit model (20) (e.g., Conti and O'Hagan, 2008; Craig et al., 2001). For Galform, we choose a product, Gaussian correlation function where each $(i, i')$ element in the $n^{(c)} \times n^{(c)}$ correlation matrix $K$, is

$$K(\theta, x_i, x_{i'}) = \prod_{l=1}^{p} \exp\Big\{ -\Big(\frac{x_{il} - x_{i'l}}{\theta}\Big)^2 \Big\}$$

and specify $\theta$ as follows. We transform the inputs so that the range of each variable is $[-1, 1]$, and consider the correlation between $\overline{\epsilon}(x_i)$ and $\overline{\epsilon}(x_{i'})$ when $|x_{il} - x_{i'l}| > 1$ for each $l$ to be practically zero (i.e., 0.00001). In turn, we solve for $\theta$ as 0.834. Given $\theta$, we use Generalized Least Squares (GLS) regression and apply a backward stepwise procedure (Venables et al., 2002) to identify the form of $g(x)$ from 967 mean model runs. We start with an overly large, full polynomial model which include the main effects, the main effects squared, the main effects cubed, and every two-way interaction between the main effects. After we reduce the model according to AIC, we result in an emulator for Galform at luminosity 18.25 (-$\mathrm{M}_{bj} + 5\log_{10}\mathrm{h}$) that has 34 coefficients, where the mean squared error = $\mathrm{Var}[\widehat{\overline{\epsilon}}(x)] = 0.013$ and adjusted $\mathrm{R}^2$=0.980.

### *9.2.3. Emulate Galform*

Given $g(x)$ and $\theta$, the computer model outcomes per dark matter specification were fit using GLS. On average the mean squared error and adjusted $\mathrm{R}^2$ were 0.00254 and 0.993 respectively. The boxplots in Figure 3 display the distribution of standardized GLS estimates for $\widehat{\beta}^{(j)}$ (centered by $\widehat{\overline{\beta}}$ and scaled by the corresponding emulator estimates for the standard errors of $\widehat{\beta}^{(j)}$). The spread for each estimate is indicative of the degree to which the system condition impacts the model outcomes.

### *9.2.4. Bayes Linear Adjustment*

Since the Galform emulator has 34 coefficients and $n^{(c)} = 967$ error terms with a common variance, we will update a $34 \times 34$ prior coefficient matrix and one univariate variance. These updates are determined by calculation (19) that relies on specifications for eight quantities: $\mu_{v_\beta}$, $\Sigma_{\mathcal{M}\{\beta\}}$, $\Sigma_{T_\beta}$, $\Sigma_{e(\widehat{\beta})}$, $\mu_{v_{\epsilon(x)}}$, $\Sigma_{\mathcal{M}\{\epsilon(x)\}}$, $\Sigma_{T_\epsilon}$, and $\Sigma_{e(\widehat{\epsilon}(x))}$. Subsequently, we assess the adjusted expectations for Galform via Equation (18) that relies on the additional specifications of $\mu_\beta$ and $\mu_\epsilon$. We advocate eliciting these values from experts by asking detailed questions about the model itself, and not the parameters specifically; the parameters are hard to conceptualize, and thus, hard to specify directly. A detailed discussion concerning the questions that we would recommend is beyond the scope of this paper. Thus, we opted to use data from a pilot study of Galform to specify the necessary prior quantities.

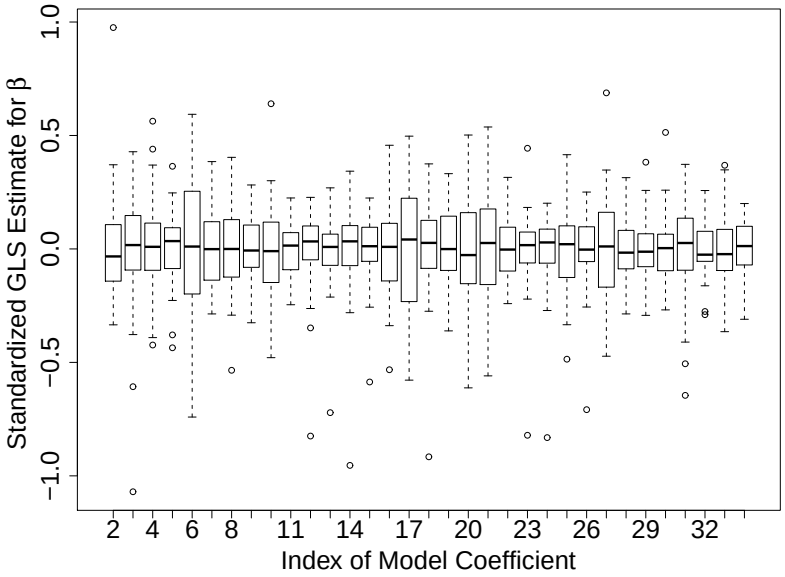



**Fig. 3.** The boxplots display the distribution of standardized GLS estimates for $\widehat{\beta}^{(j)}$ across the regions where $j \in [1, ..., m]$. Since the estimates were standardized based on $\widehat{\widehat{\beta}}$ and the corresponding region estimates for the standard errors of $\widehat{\beta}^{(j)}$, the spread for each estimate is an indication of the degree to which the system condition, $region$, impacts the model outcomes. Coefficients 2-7 represent the mains effects chosen for this emulator and represent the following: $vhotdisk$, $alpha\_reheat$, $vhotburst$, $epsilon\_Star$, $stabledisk$, and $yield$. The remaining coefficients correspond to either quadratic, cubic, or paired-interactions.

The pilot study or small computer experiment of Galform was conducted before we began our formal assessment of the latent Galform model. This experiment consisted of 100 runs for which 10 different system conditions were randomly selected per run so that, in total, 1000 runs were completed. The dataset is small and easy to manage, but unusable for the purpose of assessing the influence of inputs on process behavior because the effects of the system condition and inputs are irreversibly confounded. However, given a few guided approximations, this dataset can give reasonable prior estimates for the model quantities.

(a) We start by specifying $\mu_{v_\beta}$ and $\mu_{v_\epsilon}$. Let $\{F(x)\}_{[1000]}$ and $\overline{F}_{100}(x)$ represent respectively the sequence of model runs for the randomly selected conditions and the sequence of means for input $i$, where $i \in [1, ..., 100]$. Since the observed $100 \times 100$ variance matrix for $\{F(x)\}_{[1000]}$ estimates $\Sigma_{\mathcal{R}\{F(x)\}}$, we may solve for $\mu_{v_\beta}$ by

$$\mu_{v_\beta} = (G^T G)^{-1} G^T (\Sigma_{\mathcal{R}\{F(x)\}} - \mathrm{I}_{100}\mu_{v_\epsilon}) G (G^T G)^{-1}$$

where $G$ represents the $100 \times 34$ design matrix, $\mathrm{I}_{100}$ is a $100 \times 100$ identity matrix, and $\mu_{v_\epsilon}$ is considered to be a fraction $w_{1a}$ for the observed residual variance $\Sigma_{\mathcal{R}\{F(x)\}}$, $\mu_{v_{\epsilon(x)}} = w_{1a}\mathrm{Var}[\{F(x)\}_{[1000]}]$. We set $w_{1a} = 0.15$ because preliminary analyses of the pilot study suggested that the mean emulator, per condition, explained most of the observed output variation in the pilot study.

(b) To assess the moments of $\mathcal{M}\{\beta\}$ and $\mathcal{M}\{\epsilon(x)\}$, we rely on expert judgment and re-use the above pilot study. Namely, we set $\mu_\epsilon = \Sigma_{\mathcal{M}\{\epsilon\}} = 0$ and use the pilot data to solve for $\mu_\beta$,

$$\mu_\beta = (G^T G)^{-1} G \{\overline{F}(x)\}_{[100]}.$$

We then bootstrap the pilot data and calculate 250 estimates of $\mathcal{M}\{\beta\}$ based on samples drawn with replacement (ignoring run and region) of size 100. We set $\Sigma_{\mathcal{M}\{\beta\}}$ to the empirical variance of the 250 estimates.

(c) As explained in Section 8, we set $\Sigma_{\mathcal{M}\{v_\beta\}} = w_2\mu_{v_\beta}$, $\Sigma_{\mathcal{M}\{v_\epsilon\}} = w_2\mu_{v_\epsilon}$, $\Sigma_{\mathcal{R}\{v_\beta\}} = w_3(\Sigma_{\mathcal{M}\{v_\beta\}} + \mu^2_{v_\beta})$, and $\Sigma_{\mathcal{R}\{v_\epsilon\}} = w_3(\Sigma_{\mathcal{M}\{v_\epsilon\}} + \mu^2_{v_\epsilon})$ where $w_2$ and $w_3$ can be determined using $n^{(c)} = 967$, $\mathrm{Kur}(\beta) = \mathrm{Kur}(\epsilon) = 9$, and $p \in [0.1, .5, .9]$. For conservative reasons, we opt to assume the distributions of $\beta$ and $\epsilon(x)$ have fatter tails (comparable to those of a t-distribution with 5 degrees of freedom) than a normal distribution. To select $p$ which is the proportion of an updated variance that we expect our prior judgments to explain, we consider how well we can predict the behavior of the computer model. If we expect the observed data to alter our current beliefs dramatically, slightly, or in between, we set $p = 0.1$, $p = 0.9$, or $p = 0.5$ respectively.

(d) The final variance specifications we need are $\Sigma_{e(\widehat{\beta})}$ and $\Sigma_{e(\widehat{\epsilon}(x))}$. Since we have 967 model runs, we judge that the estimation variances should be close to zero. Thus, we use the least-squares estimates from the pilot study for the standard error of $\widehat{\overline{\beta}}$ and $\widehat{\epsilon}(x)$ and set $\Sigma_{e(\widehat{\beta})}$ to $S_\beta(X^T X)^{-1}$ and $\Sigma_{e(\widehat{\epsilon}(x))}$ to $0.05S_\epsilon$ respectively.

Following steps 1-4, we screen each prior quantity derived from the pilot study to assure that our beliefs match our specifications; e.g., to assure that we do not overstate our confidence for any given quantity by specifying a small corresponding prior variance. Once our specifications match our beliefs, we have all of the needed components to solve Equations (18) and (19) and assess $\mathrm{E}[f^{(L)}_{18.25}(x)]$ and $\mathrm{Var}[f^{(L)}_{18.25}(x)]$.

*9.2.5. Assess the Adjusted Moments for $f^{(L)}_{18.25}(x)$*

Although Bayes linear methods do not impose distributional assumptions, we display, for explanatory purposes, the results of our adjusted moments in Figure 4 as if the distribution of $f^{(L)}_{18.25}(x)$ is symmetric and compare it to the empirical distribution of realized model outcomes, $f^{(j)}_{18.25}(x)$ for $j \in [1, ..., 40]$. Specifically, we plot in Figure 4 the observed mean plus/minus three empirical standard deviations $(\overline{f}_{18.25}(x_i) \pm 3\sqrt{S_{f_{18.25}(x_i)}}$, where $S_{f_{18.25}(x_i)}$ is the sample model outcome variance at point 18.25 for run $i$), and the adjusted expected value plus/minus three adjusted standard deviations $(\mathrm{E}^{(2)}_D[f^{(L)}_{18.25}(x)] \pm 3\sqrt{\mathrm{Var}^{(2)}_D[f^{(L)}_{18.25}(x)]})$. We refer to the latter as *adjusted* intervals.

For luminosity 18.25, we do not see much difference between the observed and adjusted means, although, the empirical spread of outcomes per input tends to be smaller (with some exceptions) than the corresponding adjusted intervals. This means that the affect of dark matter specifications for some input values is greater than what we observed from the computer experiment. Relying solely on summary statistics for assessing the uncertainty of a system condition may result in overly confident model predictions. With that said, we will see in the next section that the observed variance may also over estimate model uncertainty. Underlying our calculations of the adjusted variance is a balance between the variance of $\mathcal{M}\{f(x)\}$ and the variance of $\mathcal{R}_L\{f(x)\}$. If the variance of the $\mathcal{M}\{f(x)\}$ is relatively small and the prior variance of $\mathcal{R}_L\{f(x)\}$ is less than the observed variance, the overall adjusted variance has the potential to be less than the observed variance.

To validate our methods, we implement a leave-one-out (i.e., leave-one-*region*-out) validation procedure. We re-calculated the adjusted expectation and variance forty times, each without the results (for all runs) from one of the forty regions. A high percentage of re-calculated adjusted intervals containing the missing observations would suggest that the methods proposed in this paper are reasonable. Although, 100% coverage might indicate that our adjustments are too conservative; i.e., our assessments of model uncertainty might be too large. Our cross validation procedure found that the adjusted intervals covered 99.9 percent of the missing observations.

In the next section, we extend our application to other points on the luminosity curve and repeat the cross validation procedure. We also investigate additional data that was generated by Galform conditional on regions which were not included in the original experimental design.

*9.2.6. Assess the Adjusted Moments for $f^{(L)}_{17}(x) - f^{(L)}_{22.25}(x)$*

We apply the same techniques that are described in Section 9.2.4 for prior elicitation and assess $\mathrm{E}[f^{(L)}(x)]$ and $\mathrm{Var}[f^{(L)}(x)]$ for the remaining five sensitive points on the luminosity graph (Section 9): 17, 20, 21, 21.75, and 22.25 ($-\mathrm{M}_{bj} + 5\log_{10}\mathrm{h}$). Figure (5), plots the adjusted intervals for two runs of Galform (chosen randomly from 1000): runs $x_{188}$ and $x_{870}$. Notice that our assessments of model uncertainty do not equal the observed spread of model outcomes and the difference is not uniform across the inputs nor luminosities. This is important because the behavior of Galform, or almost any complex computer model, is often understood better in some subsets of the input space and/or subsets of the responses than others. Forcing uniform estimates of uncertainty across the domain or co-domain of a computer model could hinder subsequent model-based inference. In particular, the adjusted interval at 21 ($-\mathrm{M}_{bj} + 5\log_{10}\mathrm{h}$) for run 870 is smaller than the empirical interval.

**Table 2.** Provided Galform outcomes that were based on system condition specifications (regions 41-50) that were not included in the original experimental design, this table includes the proportion of model outcomes that are contained in our adjusted intervals. Since the adjusted intervals are based only on the original data, the new predictions are similar to possible realizations of the latent model.

| New Region | Luminosity Points(- $M_{bj} + 5\log_{10}h$) | | | | | |
|---|---|---|---|---|---|---|
| | 17.00 | 18.75 | 20.00 | 21.00 | 21.75 | 22.25 |
| 41 | 1.000 | 1.000 | 1.000 | 0.996 | 0.991 | 0.995 |
| 42 | 1.000 | 1.000 | 0.998 | 0.996 | 0.989 | 0.998 |
| 43 | 1.000 | 1.000 | 0.999 | 0.997 | 0.994 | 1.000 |
| 44 | 1.000 | 1.000 | 0.999 | 0.992 | 0.988 | 0.997 |
| 45 | 1.000 | 1.000 | 0.998 | 0.994 | 0.994 | 0.997 |
| 46 | 1.000 | 1.000 | 0.999 | 0.998 | 0.993 | 0.999 |
| 47 | 0.829 | 0.997 | 1.000 | 0.995 | 0.991 | 0.999 |
| 48 | 0.863 | 1.000 | 1.000 | 0.998 | 0.995 | 0.996 |
| 49 | 1.000 | 1.000 | 0.999 | 0.992 | 0.983 | 0.997 |
| 50 | 1.000 | 1.000 | 1.000 | 0.994 | 0.989 | 1.000 |

This occurred for reasons stated in Section 9.2.5 and we observed an overly large variance due to outliers.

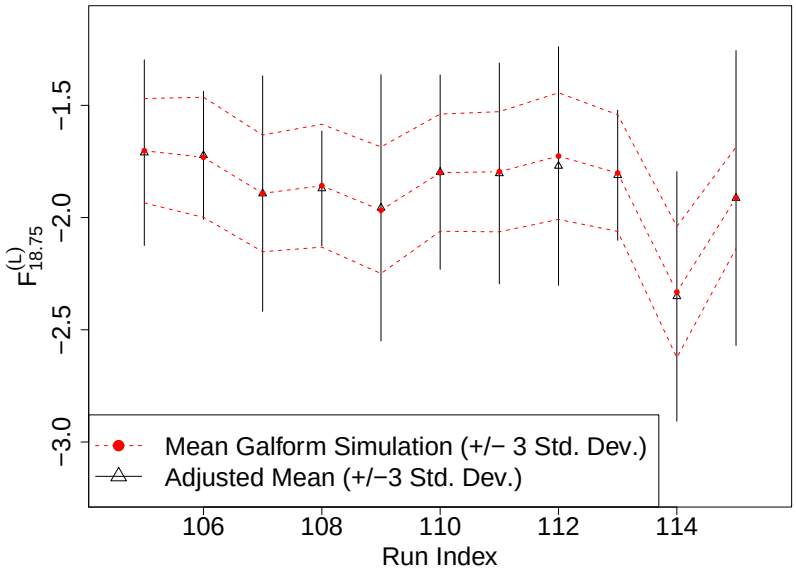
Also, the cross validation results are satisfactory and are as follows: 95.8% for outcome $f_{17}^{(L)}(x)$, 99.9% for outcome $f_{20}^{(L)}(x)$, 99.9% for outcome $f_{21}^{(L)}(x)$, 99.9% for outcome $f_{21.75}^{(L)}(x)$, and 99.9% for outcome $f_{22.25}^{(L)}(x)$. To further check our proposed methods, we requested additional Galform simulations that were based on ten *regions* that were not in the original experimental design: *regions* 41-50. We display three of the ten luminosity curves (for runs 188 and 870) that were predicted by Galform based on the new dark matter specifications in Figure 6. For each of the 10 regions and points of interest on the luminosity curve, we assess the percentage of new predictions that fall within our adjusted intervals (Table 2). Our adjusted intervals contain 83%-100% of the predictions.

## 10. Discussion

We introduced the notion of a multi-deterministic computer model which we define as a deterministic model that can produce more than one result per input because of varying system condition specifications. The main goal of this paper is to develop a method to learn about model uncertainty by pooling information across the multi-deterministic outcomes without making strong modeling assumptions, such as full exchangeability. Specifically, we show that by assuming only that the outcomes per condition are from functions that are SOEF, we can adjust prior assessments of uncertainty.

Given a multi-deterministic model, researchers often use the sample average and sample variance of the model outcomes to assess the expected value and variance of the latent model. We make two arguments against this practice. First, the mean model and the latent model differ by the latent residual term $\mathcal{R}_L\{f(x)\}$ which a posteriori need not equal zero. In fact, from the methods developed in this paper, we can learn about the discrepancy specifically. In Section 5, we defined the expectation and variance for the model results

a.




b.

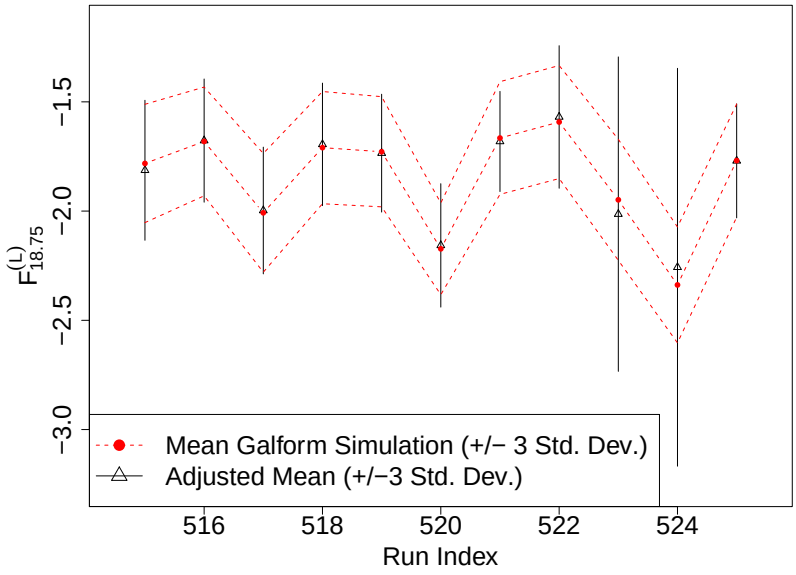
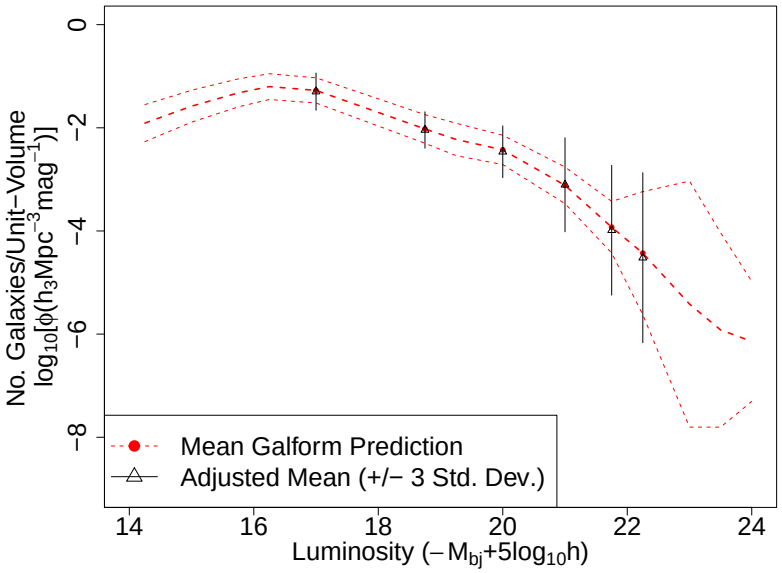



**Fig. 4.** The above graphs overlay the average simulated data and data-adjusted expectations for $f^{(L)}(x)$, $x_{\in}[x_{105}, ..., x_{115}]$ and $x_{\in}[x_{515}, ..., x_{525}]$. The circles connected by a dotted line and the parallel dotted lines represent the realized sample means $\pm 3$ standard deviations ($\overline{f}_{18.25}(x_i) \pm 3\sqrt{S_{f_{18.25}(x_i)}}$). The triangles with vertical solid lines represent the adjusted expectation $\pm 3$ adjusted standard deviations ($\mathrm{E}_D^{(2)}[f_{18.25}^{(L)}(x)] \pm 3\sqrt{\mathrm{Var}_D^{(2)}[f_{18.25}^{(L)}(x)]}$).

a.




b.

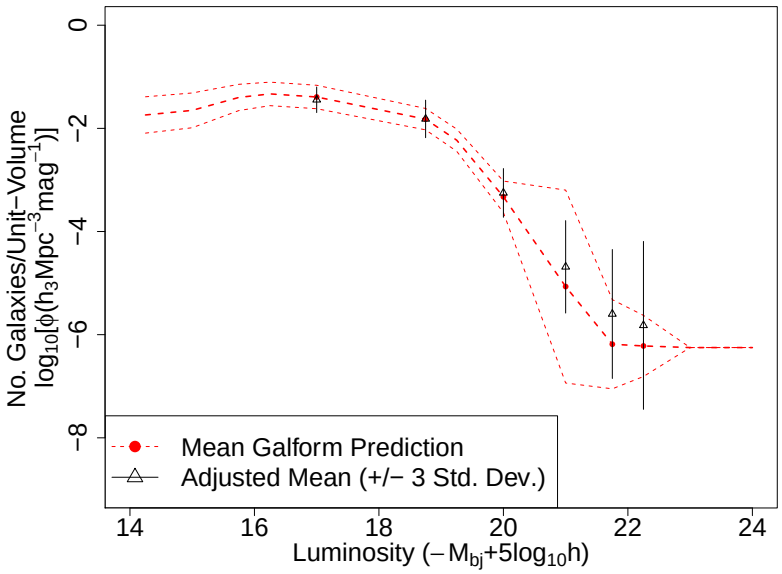
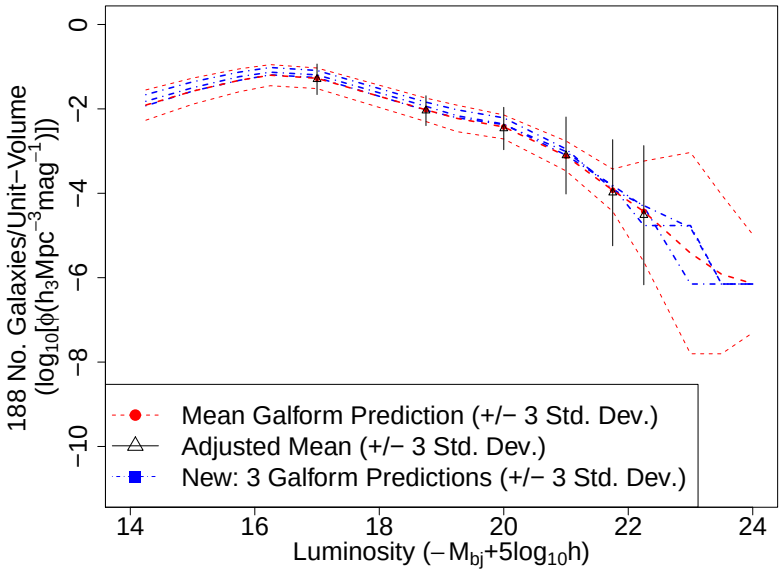



**Fig. 5.** Given analyses for six luminosity points, we may learn about the expected curve. Here, we randomly selected runs 188 and 870 to show the difference between both the simulated mean and variance across regions and our updated mean and variance. Specifically, the dotted lines represent $\pm 3$ realized, standard deviations from the realized mean curve ($\overline{f}_k(x_i) \pm 3\sqrt{S_{f_k(x_i)}}$, $k \in [17, 18.25, 20, 21, 21.75, 22.25]$). The triangles and vertical lines plot the adjusted predicted mean for a latent model and $\pm 3$ adjusted standard deviations ($\mathrm{E}_D^{(2)}[f_k^{(L)}(x_i)] \pm 3\sqrt{\mathrm{Var}_D^{(2)}[f_k^{(L)}(x_i)]}$).

a.




b.

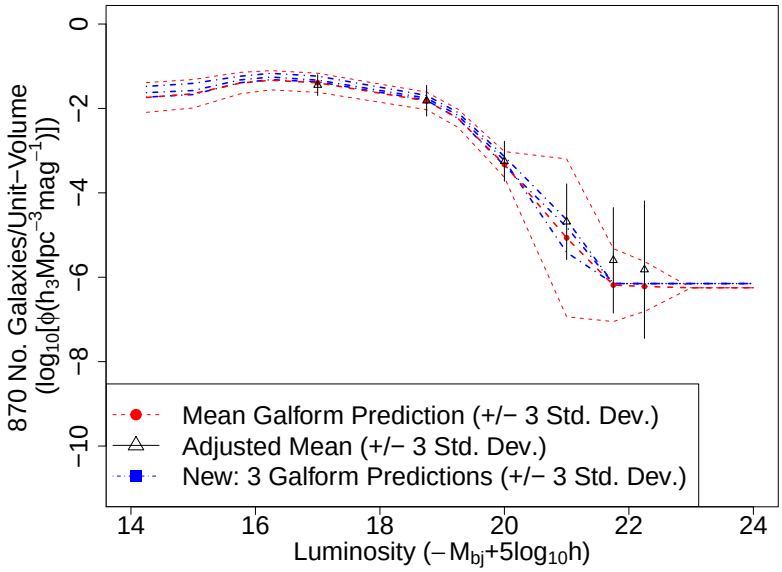



**Fig. 6.** These graphs are the same as shown in Figure 5, with three additional luminosity curves that were predicted by Galform. The predictions were based on dark matter specifications that were not included within our original experimental design and analysis. Notice that our adjusted intervals per luminosity point tend to overlap the new data and differ from the original range of outcomes.

via emulators which are a function of $\mu_\beta$, $\mu_{\epsilon(x)}$, $\Sigma_{\mathcal{M}\{\beta\}}$, $\Sigma_{\mathcal{M}\{\epsilon(x)\}}$, $\Sigma_{\mathcal{R}\{\beta\}}$, and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$. Thus, a priori,

$$\begin{aligned} \mathrm{E}[f^{(L)}(x) - f^{(c_j)}(x)] &= 0 \\ \mathrm{Var}[f^{(L)}(x) - f^{(c_j)}(x)] &= 2(g(x)\Sigma_{\mathcal{R}\{\beta\}}g(x)^T + \Sigma_{\mathcal{R}\{\epsilon(x)\}}) \end{aligned}$$

Provided posterior assessments of $\Sigma_{\mathcal{R}\{\beta\}}$, and $\Sigma_{\mathcal{R}\{\epsilon(x)\}}$ from Section 7, we obtain posterior estimates for the difference between the latent and a realized model as well. Second, approaches that rely only on summary statistics for multi-deterministic simulations may over or under estimate model uncertainty. For example, the sample variance may under estimate model uncertainty because it does not account for any variation associated with the underlying mean $\mathcal{M}\{f(x)\}$ of the computer model.

The uncertainty of the latent model associates specifically with the system condition. In this paper, we made comparisons between the affect of a system condition on computer model outcomes to the influence of a random effect in a mixed statistical model. However, we choose not to emulate the results from a multi-deterministic computer model using a mixed model for two fundamental reasons. First, a mixed statistical model requires the assumption that multi-deterministic models results per input $x$ are fully exchangeable, and we were not willing to make this strong assertion. Secondly, for moderately large $m$, $n^{(c)}$, and column-dimension $p$ of $x$, fitting a simple random intercept model with $g() = [1\ x]$ would press the memory limits in most standard computers. Hence, the methods presented in this paper do not only preserve the limits of expert judgments by relying on SOEF rather than fully exchangeable results, they also require minimal computational power to complete.

## A. Generalized Least Squares

For ease in explanation, consider $q = 1$. Let $K$ represent the correlation matrix for the computer model inputs; $F^{(j)}$ be an $n^{(c)} \times q$ matrix that contains the sequence of outcomes in matrix form for condition $j$; and $G$ denote the the matrix of inputs that are transformed according to $g(x)$, where each row $i$ equals $g(x_i)$. Given the Cholesky decomposition $K = Q^TQ$, we transform both $G$ and $F^{(j)}$ by the $Q^{-T}$ where

$$\tilde{F^{(j)}} = Q^{-T}F^{(j)} \qquad \tilde{G} = Q^{-T}G. \tag{A-1}$$

We then fit

$$f^{(c_j)}(x) = \tilde{g}(x)\beta + \epsilon(x), \quad \mathrm{Var}[\epsilon(x)] = \Sigma_\epsilon \tag{A-2}$$

based on $\tilde{F^{(j)}}$, where $\tilde{g}(x)$ equals the row of $\tilde{G}$ corresponding to $x$, $\beta^{(j)}$ is a $r \times q$ vector of model coefficients for condition $c_j$; $\epsilon^{(j)}(x)$ is a $q \times q$ matrix of error terms for condition $c_j$; $\Sigma_\epsilon$ is a $q \times q$ variance matrix for $\epsilon^{(j)}(x)$ that is independent of both $x$ and $c$.

The advantage for using GLS is that the variance matrix for all error terms is diagonal (the off-diagonal elements are zero). When $q > 1$, we may apply the same method for the vectorized version of $F^{(j)}$. In turn, $K$ is $n^{(c)}q \times n^{(c)}q$ and characterizes the correlation between model outcomes within and between runs; and, $G$ has $n^{(c)}q$ rows.

## Acknowledgements

This paper was produced with the support of the Basic Technology initiative as part of the Managing Uncertainty for Complex Models (MUCM) project, and with an EPSRC Mobility Fellowship (IV). We are also grateful to Prof Richard Bower and the rest of the Durham Semi-analytical modelling group (Institute for Computational Cosmology, Physics Department, Durham University) for providing data and their expertise in Cosmology.

## References

Carlton M. Baugh. A primer on hierarchical galaxy formation: the semi- analytical approach. *Reports on Progress in Physics*, 69:3101–3156, 2006. eprint: astro-ph/0610031.

Maria J. Bayarri, James O. Berger, Rui Paulo, Jerry Sacks, John A. Cafeo, James Cavendish, Chin-Hsu Lin, and Jian Tu. A Framework for Validation of Computer Models. *Technometrics*, 49(2):138–154, 2007.

R. G. Bower, A. J. Benson, R. Malbon, J.C. Helly, C. S. Frenk, C.M. Baugh, S. Cole, and C. G. Lacey. The broken hierarchy of galaxy formation. *Monthly Notices of the Astronomical Society*, (370):645–655, 2006.

Shaun Cole, Alfonso Aragon-Salamanca, Carlos S Frenk, Julio F Navarro, and Stephen E Zepf. A recipe for galaxy formation. *Royal Astronomical Society, Monthly Notices*, 271 (4):781–806, 15 Dec. 1994 1994.

Stefano Conti and Athonay O'Hagan. Bayesian emulation of complex multi-output and dynamic computer models. Technical report, University of Sheffield, Department of Probability and Statistics, The Hicks Building, University of Sheffield, Sheffield S3 7RH, UK, 2008.

Peter S Craig, Michael Goldstein, Jonathan C Rougier, and Allan H Seheult. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, 96(454):717–729, 2001.

Bruno de Finetti. *Theory of Probability*, volume 1. New York: John Wiley and Sons, 1974.

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

Michael Goldstein and Jonathan Rougier. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, 101(475):1132–1143, 2006a.

Michael Goldstein and Jonathon Rougier. Reified bayesian modeling and inference for physical systems. *Journal of Statistical Planning and Inference*, 2006b.

Michael Goldstein and David Woof. *Linear Bayes - FILLIN*. JOe SMITH, 2007.

Michael Goldstein, Leanna House, and Jonathon Rougier. Analysing model discrepancy from observations in a multimodel ensemble. Technical report, Bristol, 2008.

Marc C. Kennedy, Clive W. Anderson, Conti Stefano, and Anthony O'Hagan. Case Studies in Gaussian Process Modelling of Computer Codes. *Reliability Engineering and System Safety*, 91:1301–1309, 2006.

Charles E. McCulloch. Generalized linear mixed models. In *Encyclopedia of Environmetrics*, pages 869–873. John Wiley & Sons, 2002.

Anthony O'Hagan. Eliciting expert beliefs in substantial practical applications. *The Statistician*, 47(1):21–35, 1998.

Jonathon Rougier. Efficient emulators for multivariate deterministic functions. Technical report, University of Bristol, Department of Mathematics, Univeristy Walk, Bristol BS8 1TW, UK, 2008.

J.O. Sewall, R. S. W. van de Wal, K. van der Zwan, C. van Oosterhout, H. A. Dijkstra, and C. R. Scotese. Climate model boundary conditions for four cretaceous time slices. *Climate of the Past*, 3:647–657, 2007.

Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, John A. Peacock, Shaun Cole, Peter Thomas, Hugh Couchman, August Evrard, Joerg Colberg, and Frazer Pearce. Simulating the joint evolution of quasars, galaxies and their large-scale distribution. *Nature*, 435:629–636, 2005.

W. N. Venables, Brian D. Ripley, and W. N. Venables. *Modern Applied Statistics with S*. Springer-Verlag Inc, 2002. ISBN 0-387-95457-0.