Learning about physical parameters: The importance of model discrepancy

Jenný Brynjarsdóttir[1] and Anthony O'Hagan[2]

June 9, 2014

Abstract

Science-based simulation models are widely used to predict the behavior of complex physical systems. It is also common to use observations of the physical system to solve the inverse problem, i.e. to learn about the values of parameters within the model, a process often called *calibration*. The main goal of calibration is usually to improve the predictive performance of the simulator but the values of the parameters in the model may also be of intrinsic scientific interest in their own right.

In order to make appropriate use of observations of the physical system it is important to recognise *model discrepancy*, the difference between reality and the simulator output. We illustrate through a simple example that an analysis that does not account for model discrepancy may lead to biased and over-confident parameter estimates and predictions.

The challenge with incorporating model discrepancy in statistical inverse problems is the confounding with calibration parameters, which will only be resolved with meaningful priors. For our simple example, we model the model-discrepancy via a Gaussian Process and demonstrate that by accounting for model discrepancy our prediction within the range of data is correct. However, only with realistic priors on the model discrepancy do we uncover the true parameter values. Through theoretical arguments we show that these findings are typical of the general problem of learning about physical parameters and the underlying physical system using science-based mechanistic models.

Keywords: Computer models, Uncertainty Quantification, Model inadequacy, Model error, Model form error, Model bias, Structural uncertainty, Calibration, Extrapolation, Simulation model.

[1]Case Western Reserve University, Cleveland, Ohio (*email:* jenny.brynjarsdottir@case.edu)
[2]The University of Sheffield, United Kingdom (*email:* a.ohagan@sheffield.ac.uk)

1 Introduction

Simulation models are used across almost all areas of physical science, engineering, environmental, social and health sciences to study, predict and control the behavior of complex physical systems. Such models are generally based on current scientific understanding, often involving large systems of differential equations and implemented in complex computer programs. We call these *simulators* and here we are concerned with *deterministic* simulators, i.e. ones that do not inherently model stochasticity and there exists a single true underlying solution. Deterministic simulators are increasingly being employed in decision-making at all levels from individual companies through national governments to international agencies, and as a result there is a growing emphasis on understanding the uncertainties in their outputs.

The field of study that has in recent years acquired the name 'Uncertainty Quantification', or simply 'UQ', has primarily been concerned with propagating uncertainty about the inputs in order to infer the uncertainty in the simulator output. However, another important component of uncertainty is *model discrepancy*, defined as the difference between the simulator output and the real physical system value that the simulator is intended to predict. Simply propagating input uncertainty to infer output uncertainty is clearly only a partial quantification of uncertainty about the physical system value as it does not take into account model discrepancy (and other sources of uncertainty).

Model discrepancy was formally introduced as a source of uncertainty in simulator predictions by Kennedy and O'Hagan (2001), who referred to it as model inadequacy (other commonly used names include model error, model form error, model bias and structural uncertainty). They considered the problem of using observations of the real physical system to learn about uncertain input parameters, a process usually known as calibration, and showed how to account for model discrepancy in calibration and in subsequent predictions of the physical system. Since then, their inferential ideas and modeling framework have been widely adopted and further developed, e.g. Higdon et al. (2004, 2008); Qian and Wu (2008); Goldstein and Rougier (2009); Bayarri et al. (2007); Gramacy and Lee (2008).

Furthermore, the sometimes called "Kennedy-and-O'Hagan approach" has been applied in diverse fields such as ecology (Arhonditsis et al., 2008), hydrology (Reichert and Mieleitner, 2009), engineering (Apley et al., 2006; Bayarri et al., 2009), health sciences (Strong et al., 2012), nuclear reactors (Unal et al., 2011), experimental physics (Lee et al., 2008), astrophysics (Habib et al., 2007) and climate modeling (Murphy et al., 2007; Stainforth et al., 2007; Rougier, 2007; Sansó and Forest, 2009).

Modeling of model discrepancy uncertainty by Kennedy and O'Hagan (2001) was simple but they claimed that by recognising this additional source of uncertainty the resulting predictions were not over-confident. We consider here the role of model discrepancy in more detail, using general arguments and supporting those arguments with a simple example. We show that while the original claim of Kennedy and O'Hagan (2001) may be correct when we consider predictions in the region where we have data, i.e. interpolation, the situation is more complex when it comes to learning about the true values of physical parameters in the model and when our interest is in extrapolation.

Our principal findings concerning the use of observations of the physical system to learn about parameters or to make predictions are:

- if model discrepancy is ignored, predictions (both interpolations and extrapolations) and inferences about parameters are biased, and this bias persists with increasing numbers of observations;
- if model discrepancy is modelled in a simple, uninformative way, interpolations are unbiased but extrapolations and inferences about parameters will still typically be biased;
- in order to obtain realistic learning about model parameters, or to extrapolate outside the range of the observations, it is important not just to incorporate model discrepancy but to model carefully the available prior information about it.

We do not claim that these findings are entirely novel; certainly many researchers working with the Kennedy-and-O'Hagan approach are aware of some or all of these results. Nevertheless, there has not previously been a full exposition of these findings in the statistics literature, let alone in applied mathematics journals. Our claim to originality lies in the careful and thorough exposition of the importance of model discrepancy and of prior information about model discrepancy, aided by a simple and intuitive example.

We begin in Section 2 by formulating the calibration problem in general terms. Section 3 introduces a simple example to motivate the principal findings above, then in Section 4 we present the general theory. In Section 5 we summarise the findings and argue that the development of techniques for modelling prior knowledge about model discrepancy is possibly the highest priority topic for research in UQ today.

2 Calibration

2.1 Notation

We formulate calibration using the terminology and notation of Kennedy and O'Hagan (2001). A simulator will typically have some inputs, known as *control variables*, whose values define the particular instance of the physical system to be predicted, e.g. location, time, machine setting etc. It also has inputs known as *calibration parameters* whose values are uncertain. We represent the simulator as $\eta(x, u)$, where x denotes the control variables and u represents the calibration parameters. We suppose that we have n observations of the physical system, denoted by $z_1, z_2, \ldots, z_n$. The i-th observation z_i is of an instance of the physical system associated with control inputs x_i, and we assume that the x_is are known.

We model the i-th observation as

$$z_i = \zeta(x_i) + \epsilon_i \ , \tag{1}$$

where $\zeta(x_i)$ is the true value of the physical system at control variable value x_i and the ϵ_is are independent observation errors. Model discrepancy now enters through the equation linking reality to the simulator:

$$\zeta(x) = \eta(x, \theta) + \delta(x) \ , \tag{2}$$

where θ is the true but unknown value of the calibration parameter vector u. Combining equations (1) and (2), we have

$$z_i = \eta(x_i, \theta) + \delta(x_i) + \epsilon_i \ , \tag{3}$$

which expresses the observation as the sum of three terms — the simulator output evaluated at the relevant values x_i of the control variables and the true values θ of the calibration parameters, the model discrepancy at x_i and observation error.

We take a Bayesian approach, that is, prior distributions are assigned to θ and the model discrepancy function $\delta(\cdot)$, and these are updated to posterior distributions conditional on the observations. A posterior predictive distribution for the value of the physical system $\zeta(x)$ at any given x can then be derived from (2). We will show that if model discrepancy is ignored then the posterior distributions of θ and $\zeta(x)$ will not have desirable properties, and in particular will not converge on the true values as the sample size n goes to infinity.

2.2 Physical parameters and tuning parameters

It is sometimes useful to identify two types of calibration parameters. Physical parameters have meaning within the science underlying the simulation model. Learning about the values of physical parameters can contribute to the underlying science. Tuning parameters, however, do not have physical meaning. Their role is often to act as a simplified surrogate for some more complex process that is not modelled in the simulator. They are there to help the model to predict reality more accurately and their 'true' values are whatever enable the model to fit best to reality. Although learning about the values of tuning parameters may have limited scientific value, for instance to learn something about the practical impact of processes not directly modelled or resolved in the simulator, they are not of scientific interest in their own right.

Physical parameters are also important for extrapolation. Simulators are typically calibrated using observations at x_i values for which it is feasible to make such observations, but the simulator is generally required to predict the behaviour of the physical system at control variable values where we cannot readily observe it. Tuning parameters do not help to make the leap from where we have data to where we do not, because the best fitting values of tuning parameters will typically be different. Because physical parameters have fixed values independent of the context of the simulator's application, learning about their values with observations at some x_is should improve prediction at other values of x.

3 A motivating example

We present the following analysis of a simple example to introduce and motivate the general analysis in Section 4.

3.1 The simple machine

Imagine a machine where the amount of work delivered depends on the amount of effort we put into it. Our somewhat naive simulator for this machine is that work is proportional to effort, i.e.

$$\eta(x, \theta) = \theta x \tag{4}$$

where x is effort and θ is the efficiency of the machine. The parameter θ is unknown and our interest is in both learning about θ and then using this model to predict work for a given amount of effort. In the notation of Section 2, x is a control input, θ is a physically

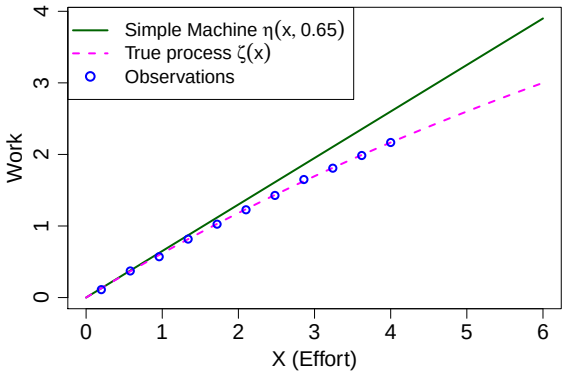

Figure 1: The Simple Machine for the true value of θ: $\eta(x, 0.65) = 0.65x$, the true process $\zeta(x)$ and the dataset with 11 observations.

meaningful calibration parameter and there are no tuning parameters. This simulator has obvious deficiencies in that any losses in efficiency (e.g. due to friction) are not accounted for. We therefore have *model-discrepancy* that needs to be acknowledged and dealt with.

We produced synthetic experimental data by simulating, with random error, from the true process;

$$z_i = \zeta(x_i) + \epsilon_i \,, \qquad i = 1, \ldots, n \tag{5}$$

where the measurement errors ϵ_i are i.i.d. $N(0, 0.01^2)$ and the inputs $x_1, \ldots, x_n$ were evenly spaced over the interval $[0.2, 4]$. The underlying true process is

$$\zeta(x) = \frac{\theta x}{1 + x/a} \tag{6}$$

where $\theta = 0.65$ and $a = 20$. To illustrate the effect of sample size we produced three datasets that have $n = 11$, 31 and 61, where the smaller datasets are subsets of the larger datasets. Figure 1 shows the Simple Machine for the true value of θ, i.e. $\eta(x, 0.65)$, the true process $\zeta(x)$ and the dataset with eleven observations.

Our computer model is simple, yet it is a sensible model for the process. The θ parameter in the Simple Machine (4) is the same physical parameter as in true process (6). That is, for small x the true process is roughly θx and θ is the gradient of the true process at zero. It is the theoretical efficiency of the machine, and as such it is a physical parameter that is of intrinsic interest to the machine's designers. From now, we will act as though we do not know the true process in (6) exactly. We only use the synthetic datasets and the more vague information about the discrepancy between $\zeta(x)$ and $\eta(x, \theta)$ that "losses are not accounted for".

3.2 Analysis without accounting for model discrepancy

For the Simple Machine, if we ignore the model discrepancy term in (3) we have

$$z_i = \eta(x_i, \theta) + \epsilon_i = \theta x_i + \epsilon_i \ , \quad i = 1, \ldots, n \ , \tag{7}$$

where ϵ_i are i.i.d. $N(0, \sigma_\epsilon^2)$. This is just a linear regression through the origin, and we estimate θ using the usual method of Bayesian regression. We assume the joint improper prior distribution $p(\theta, \sigma_\epsilon^2) \propto \sigma_\epsilon^{-2}$ and so posterior inference on θ is based on the Student's t_{n-1} distribution (see Appendix A).

The posterior densities of θ for the three different datasets are shown in Figure 2 a). The posterior means and 90% credible intervals are

$$\begin{aligned} &\text{11 observations:} \quad 0.562 \quad (0.551, 0.573), \\ &\text{31 observations:} \quad 0.564 \quad (0.557, 0.571) \text{ and} \\ &\text{61 observations:} \quad 0.565 \quad (0.560, 0.569) \ . \end{aligned}$$

In this simple example, it is easy to see why we under-estimate θ. The straight line through the origin that best fits the data in Figure 1 will clearly have a slope less than the true $\theta = 0.65$, which is the gradient at the origin. It is important to recognise that the bias, which will clearly persist for any number of observations, is due to model discrepancy.

Not only is θ under-estimated but the posterior credible intervals are not even close to covering the true parameter value. This is in part because we have fixed the observation error variance σ_ϵ^2 at a small value, so the posterior variance of θ is quite small even with only 11 observations. However, even if we had used a larger error variance, with more data the posterior variance will become arbitrarily small and so credible intervals will always fail to cover the true value for sufficiently large n. Quite simply, with more and more data from the true process we become more and more sure about the wrong value for θ.

Figures 3 a) and 5 a) show the posterior density of $\zeta(x_0)$, which since we are ignoring model discrepancy is the same as θx_0, at $x_0 = 1.5$ (interpolation) and $x_0 = 6$ (extrapolation) for the three sample sizes. The corresponding true values are $\zeta(1.5) = 0.65 \times 1.5/(1+1.5/20) = 0.907$ and $\zeta(6) = 3.0$. In every case the posterior density fails to cover the true value (even for interpolation) and with more data we only become more and more sure about the wrong value.

3.3 Accounting for model discrepancy — methods

We present two analyses that account for model discrepancy but differ in the prior that is assumed for the discrepancy term.

For our Simple Machine the observation equation (3) becomes

$$z_i = x_i\theta + \delta(x_i) + \epsilon_i \ , \quad i = 1, \ldots, n \ , \tag{8}$$

where the ϵ_is are independent $N(0, \sigma_\epsilon^2)$ and we now explicitly acknowledge model discrepancy. Following Kennedy and O'Hagan (2001), we represent the model discrepancy term $\delta(x)$ as a zero-mean Gaussian process (GP):

$$\delta(\cdot) \sim GP(0, \sigma^2 c(\cdot, \cdot|\psi)) \ , \tag{9}$$

with the squared exponential correlation function (also called the Gaussian correlation function):

$$c(x_1, x_2|\psi) = \exp\left(-\left(\frac{x_1 - x_2}{\psi}\right)^2\right) \ . \tag{10}$$

It is important to understand how such a representation might formulate prior knowledge about the discrepancy function δ. A GP is a probability distribution for a function. Equation (9) says that at any point x the prior probability distribution of $\delta(x)$ is normal (Gaussian) with mean zero and variance $\sigma^2 c(x, x \mid \psi)$, which from (10) is just σ^2. The zero mean says that we do not have a prior expectation that $\delta(x)$ is more likely to be positive or more likely to be negative. The variance σ^2 expresses a prior belief that $\delta(x)$ is not likely to be outside the range $\pm 2\sigma$, so it measures the strength of prior information about $\delta(x)$. The fact that the variance is the same for all x implies that we do not have a prior expectation that $|\delta(x)|$ is likely to take larger values for some x values than for others. The correlation function (10) expresses a prior belief that $\delta(x)$ will be a smooth function, with the value of $\delta(x_1)$ being close to that of $\delta(x_2)$ if x_1 is close to x_2. The parameter ψ determines how far apart x_1 and x_2 need to be before $\delta(x_1)$ can be very different from $\delta(x_2)$.

Our first analysis using model discrepancy adopts this simple representation, as originally proposed by Kennedy and O'Hagan (2001). Figure 6 (first column) shows a few realizations of this prior. We know in advance that this is not a realistic prior for the model discrepancy in the Simple Machine example because it goes against some of the things we know a priori. We know that our Simple Machine does not account for friction so the real process will be less than what our Simple Machine predicts (using the true θ). Hence, we know that $\delta(x)$ is always negative (except at $x = 0$) which is not reflected in the zero-mean GP prior. Also, we know that the friction will increase slowly and smoothly as the effort x increases so $\delta(x)$

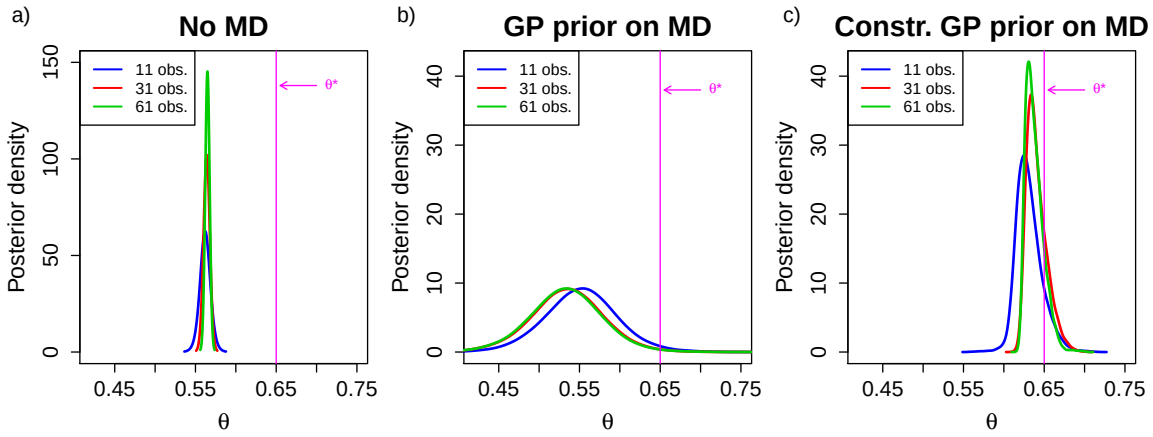

Figure 2: Posterior densities of θ for the three cases: a) Analysis without model-discrepancy (MD), b) assuming a Gaussian Process (GP) prior on the MD and c) assuming a constrained GP prior on the MD. The true value of θ (0.65) is indicated with a vertical line.

is a smooth and monotonically decreasing function. At the other end, friction will tend to zero when x tends to zero and in fact we know that $\delta(0) = 0$. Any serious analysis would respect these properties in a prior distribution for the model discrepancy.

As a second approach we therefore incorporate our prior knowledge about the model discrepancy by conditioning the process and its derivatives at pre-specified points. What makes this possible is that derivatives of a Gaussian process are still a Gaussian process. More specifically, if $\delta(x)$ is a GP with mean function $m(x)$ and covariance function $\sigma^2 c(x_1, x_2)$ then all derivatives $\delta^j(x) = \frac{d^j \delta(x)}{dx^j}$ are jointly normal with

$$E(\delta^j(x)) = \frac{d^j m(x)}{dx^j} \quad \text{and}$$
$$\mathrm{Cov}(\delta^j(x_1), \delta^i(x_2)) = \sigma^2 \frac{d^{j+i} c(x_1, x_2)}{dx_1^j dx_2^i}$$

given that the above derivatives exist (O'Hagan, 1992; Adler, 2010). Therefore, after conditioning on specific values of the process and/or its derivatives at fixed points we still have joint normality. Here we condition on

$$\delta(0) = 0 \quad \text{and} \quad \delta'(0) = 0\,. \tag{11}$$

These constraints reflect that $\delta(x)$ tends to zero when x tends to zero as well as being zero at the origin. The constraint that $\delta'(0) = 0$ is very important because it is through this condition that we express the true physical meaning of θ as the slope of $\zeta(x)$ at $x = 0$. We also know that $\delta(x) \leq 0$ and $\delta'(x) \leq 0$ (model-discrepancy is monotone) for all x. We cannot impose such global inequality constraints and still enjoy the computational convenience of

a Gaussian Process. However we can introduce inequality constraints at a finite number of points which leads to (multivariate) truncated normal distributions. Examples of imposing constraints on the derivatives of a Gaussian Process include Riihimäki and Vehtari (2010) and Wang and Berger (2011). More discussions on this topic can for example be found in Da Veiga and Marrel (2012) and references therein.

Figure 6 (second column) shows realizations of a Gaussian process conditioned on the equality constraints in (11) and two inequality constraints

$$\delta'(0.5) < 0 \quad \text{and} \quad \delta'(1.5) < 0\,. \tag{12}$$

We see how the effects of the locations of the inequality constraints and the correlation length ψ interact. If the process is smooth (ψ is large) then only a few constraints of the form $\delta'(x_i) < 0$ are needed to keep the realizations negative. If ψ is small the realizations are free to wander off between the locations of the inequality constraints. In our case we know that the model discrepancy is smooth, which will be reflected in our prior on the correlation length ψ, so it should be enough to include only a few inequality constraints to impose monotonicity.

We fit the model in (8) for these two cases, the unconstrained case and the constrained case, applying the four constraints in (11) and (12). We assume the same prior distributions for unknown parameters in both cases. We assume a flat improper prior for θ, imposing no prior knowledge on θ. In practice there may be some prior knowledge about the physical parameters (range, order of magnitude), but our focus here is on learning θ and on priors for model discrepancy. We use Inverse-Gamma priors for σ_ϵ^2 and σ^2. We assume that the measurement error is fairly well known and choose a mean of 0.01^2 and mode 0.009^2 for the prior of σ_ϵ^2 (recall that the synthetic data were simulated with $\sigma_\epsilon^2 = 0.01^2$). Less is known about σ^2 and we choose the Inverse-Gamma distribution with mean 0.3^2 and mode 0.2^2 based on the order of magnitude of the work produced. We know that the model discrepancy is smooth and with that in mind we assume a Gamma prior distribution for ψ with mean 1 and variance 0.2. Due to numerical problems in the MCMC algorithm we truncate this prior distribution at $\psi = 4$.

In both cases we obtain approximate samples from the posterior distribution via Gibbs sampling with a Metropolis-Hastings step for ψ. Full conditional distributions are given in Appendix B. Since we are interested in inference about the shape of the model discrepancy function we sample the $\delta(x)$ process at points of interest as well as other parameters. One problem that arises is that when ψ is large the covariance matrices become numerically singular which creates problems for the Gibbs sampler. We remedy that by sampling $\delta(x_i)$ only at 6 observation points, $\boldsymbol{\delta}_o$, and set the rest of $\delta(x_j)$'s equal to their conditional

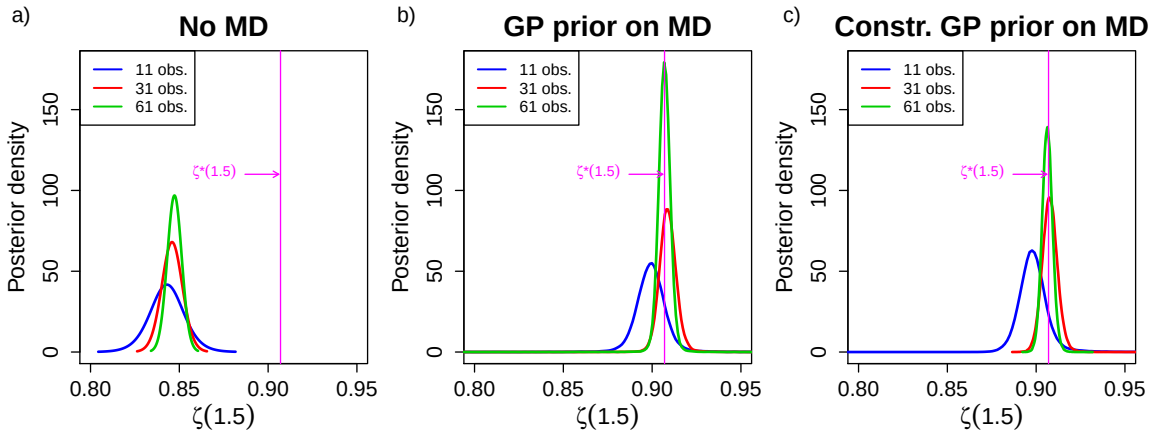

Figure 3: Posterior densities of the true process at $x_0 = 1.5$ (interpolation) for the three cases: a) Analysis without model-discrepancy (MD), b) assuming a Gaussian Process (GP) prior on the MD and c) assuming a constrained GP prior on the MD. The true value of $\zeta(1.5)$ is indicated with a vertical line.

mean, given $\boldsymbol{\delta}_o$. The details are given in Appendix B. For the unconstrained case we obtained 80,000 samples from the posterior and discarded the first 20,000 as burn-in. In the constrained case the parameters showed slower mixing so we obtained 200,000 samples and discarded the first 50,000 as burn-in. Traceplots of parameters did not indicate lack of convergence.

3.4 Analysis with model discrepancy — results

In Figure 2 (second two columns) we show estimated posterior densities of θ for the three datasets and both the constrained and unconstrained prior distributions for $\delta(\cdot)$. The unconstrained GP prior on the discrepancy yields the following posterior means and 90% credible intervals:

$$\begin{aligned} &\text{11 observations:} \quad 0.552 \quad (0.477, 0.626), \\ &\text{31 observations:} \quad 0.535 \quad (0.459, 0.609) \text{ and} \\ &\text{61 observations:} \quad 0.533 \quad (0.458, 0.607)\ . \end{aligned}$$

As in the analysis without model discrepancy, we see that the posterior distribution is biased downwards; with increasing sample size the posterior mean is too low and the true physical value of 0.65 is always out in the extreme tail of the distribution. However, unlike the previous analysis the posterior uncertainty does not degenerate with more data. The width of the credible interval does not tend to zero but seems to have stabilised by around $n = 31$. Turning to the posterior densities from the constrained GP prior (see last column

Unconstrained prior on the model discrepancy

	ψ		σ		σ_ϵ	
11 obs.	1.94	(1.08, 2.89)	0.20	(0.13, 0.30)	0.0111	(0.0088, 0.0142)
31 obs.	2.11	(1.34, 2.96)	0.20	(0.14, 0.30)	0.0120	(0.0100, 0.0143)
61 obs.	2.14	(1.46, 2.89)	0.21	(0.14, 0.31)	0.0114	(0.0099, 0.0131)

Constrained prior on the model discrepancy

	ψ		σ		σ_ϵ	
11 obs.	2.02	(1.13, 2.95)	0.25	(0.16, 0.38)	0.0110	(0.0087, 0.0139)
31 obs.	2.22	(1.47, 3.11)	0.27	(0.18, 0.42)	0.0119	(0.0100, 0.0142)
61 obs.	2.23	(1.59, 3.02)	0.28	(0.18, 0.42)	0.0114	(0.0100, 0.0130)

Table 1: Posterior means and 90% credible intervals for ψ, σ^2 and σ_ϵ^2 for different cases.

in Figure 2) we see that these do cover the true value of θ although concentrated slightly to the left of it. The posterior means and 90% credible intervals are

$$\begin{aligned} &\text{11 observations: } 0.631 \ (0.610, 0.661), \\ &\text{31 observations: } 0.641 \ (0.624, 0.665) \text{ and} \\ &\text{61 observations: } 0.638 \ (0.624, 0.659) \ . \end{aligned}$$

Like the analysis with the unconstrained prior, with increasing n the width of the credible interval stabilises without tending to zero, although they are now much narrower because of the extra information that the constraints imply about $\delta(\cdot)$.

Posterior means and 90% credible intervals for the range parameter, ψ, and the standard deviations σ and σ_ϵ are shown in Table 1 and posterior densities of ψ are shown in Figure 4. The posterior inference of these parameters are very similar for the unconstrained and constrained cases.

Now consider inference about the true process $\zeta(x_0)$ at an x_0 where it has not been observed. In Section 3.2 we saw that without model discrepancy the posterior distributions failed to cover the true value, both for interpolation at $x_0 = 1.5$ and for extrapolation at $x_0 = 6$. Estimated posterior predictive densities for $\zeta(1.5)$ are shown in Figure 3. Compared with the case without model discrepancy in Figure 3 panel a), panels b) and c) show that accounting for model-discrepancy leads to a dramatic improvement. The posterior predictive densities not only cover the true value of $\zeta(1.5) = 0.907$ they become centered around the true value when sample size increases. However, introducing model discrepancy is far less effective when we consider extrapolation. Figure 5 shows estimated predictive posterior densities of $\zeta(6)$. In all cases we completely over shoot the true value and only in the unconstrained case do the densities reach the true value.

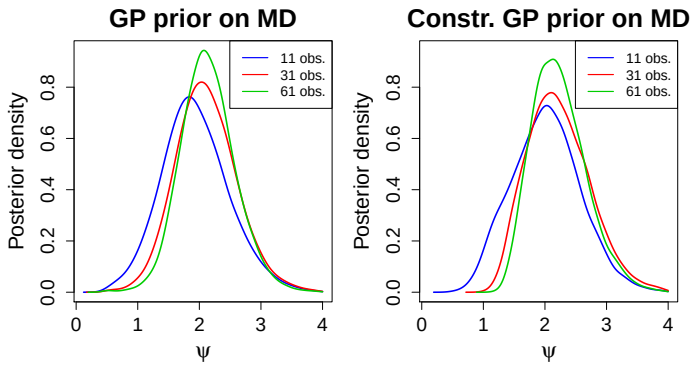

Figure 4: Posterior densities of the range parameter, ψ, of the Gaussian Process (GP) modeling the model-discrepancy (left) and the constrained GP (right).

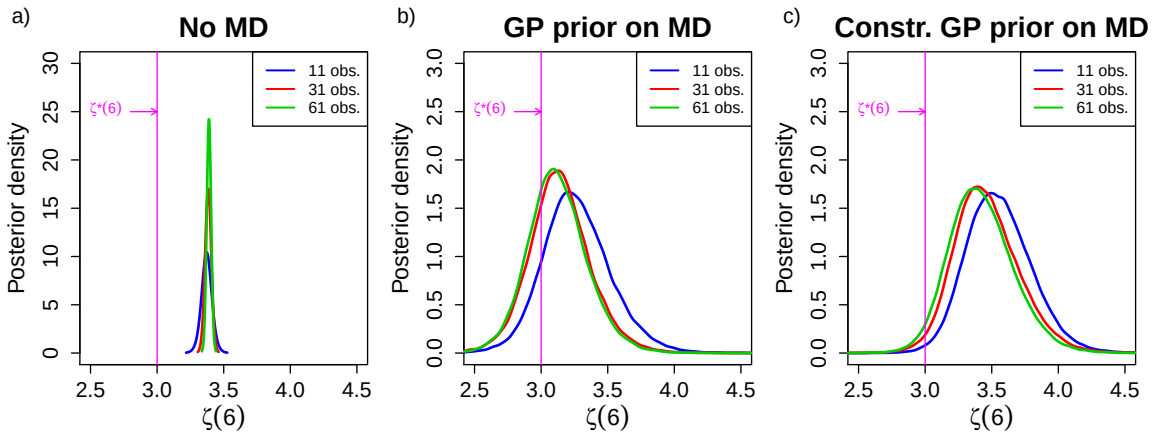

Figure 5: Posterior densities of the true process at $x_0 = 6$ (extrapolation) for the three cases: a) Analysis without model-discrepancy (MD), b) assuming a Gaussian Process (GP) prior on the MD and c) assuming a constrained GP prior on the MD. The true value of $\zeta(6)$ is indicated with a vertical line.

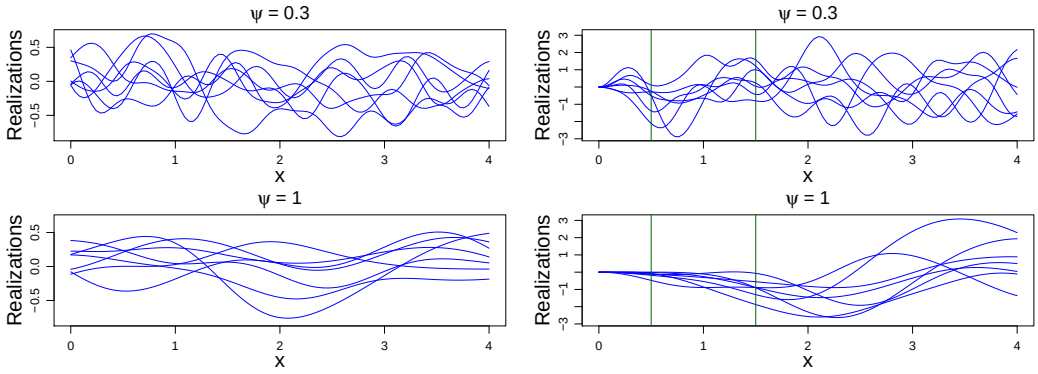

Figure 6: Realizations of a Gaussian process with mean function zero, covariance function squared exponential, $\sigma^2 = 1$ and correlation length $\psi = 0.3$ (first row) or $\psi = 1$ (second row). First column is the unconstrained process, in the second column we have constrains $\delta(0) = 0$, $\delta'(0) = 0$, $\delta(0.5) < 0$ and $\delta(1.5) < 0$.

Introducing model discrepancy has clearly had mixed results for the Simple Machine and can be summarized as follows:

- It seems that in order to make good inferences about the physical parameter θ we need to do more than just introduce the model discrepancy term. In this example, it is only when we add constraints which represent genuine prior information about model discrepancy that the posterior credible intervals cover the true value.
- When we look at interpolation, adding model discrepancy in this example produces good inferences, even with the unconstrained GP prior.
- For extrapolation, introducing model discrepancy does not appear to have made good inferences possible. Furthermore, in this example it seems that introducing constraints on $\delta(x)$ in the range of the data does not help (and may even make inferences worse than with the simple unconstrained prior) when extrapolating outside that range.

In the next section we explain these findings and show that they are typical of what we should expect to find in general when trying to use science-based mechanistic models to learn about physical parameters and the underlying physical system.

4 The general case

4.1 Calibration without model discrepancy

Consider the general model (3) for observations, where model discrepancy $\delta(x_i)$ is included explicitly. We could write it as

$$z_i = \eta(x_i, \theta) + e_i \ , \tag{13}$$

where model discrepancy has now been absorbed into the new error term $e_i = \delta(x_i) + \epsilon_i$. However, it is important to recognise that whereas the original observation errors ϵ_i are independent this will not be true of the e_is because it would require the discrepancy terms $\delta(x_i)$ and $\delta(x_j)$ to be independent for all $i \neq j$. Independence clearly cannot hold if the i-th and j-th observations are made at the same control values, because then $\delta(x_i)$ must equal $\delta(x_j)$. Furthermore, model discrepancy will generally be a continuous function of x, and therefore $\delta(x_i)$ and $\delta(x_j)$ will be correlated whenever x_i is sufficiently close to x_j.

The traditional approach to calibration ignores model discrepancy. As in Section 3.2 for the Simple Machine example, it is usual to assume (13) but to treat the e_is as simple independent observation errors ϵ_i. That is, traditional calibration uses the model

$$z_i = \eta(x_i, \theta) + \epsilon_i \ , \tag{14}$$

and generally assumes that the ϵ_is are independent $N(0, \sigma_\epsilon^2)$ errors. In Section 3.2 this reduced to the linear regression model (7). In general, (14) is a nonlinear regression model. With weak prior information, the posterior distribution of θ will be centred on the best fitting value $\hat{\theta}$ which minimises the residual mean square

$$S(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ z_i - \eta(x_i, \hat{\theta}) \right\}^2 \ .$$

Because the correct representation of the z_is is (3), this estimate will be biased, even if the sample size is large.

To understand this effect it is helpful to consider the limit as the number of observations goes to infinity. Suppose, therefore, that we have effectively an infinite number of observations with x_i values filling some subspace $\mathcal{X}_{obs}$ of the space of possible values of x. Then in the limit $S(\hat{\theta})$ will become $\sigma_\epsilon^2 + S_\infty(\hat{\theta})$, where

$$S_\infty(\hat{\theta}) = \int_{\mathcal{X}_{obs}} \left\{ \zeta(x) - \eta(x, \hat{\theta}) \right\}^2 p(x) dx \ , \tag{15}$$

where $p(x)$ denotes the density of the observations over $\mathcal{X}_{obs}$. The posterior mean of θ will be the value of $\hat{\theta}$ which minimises $S_\infty(\hat{\theta})$. If we substitute the true value $\hat{\theta} = \theta$ into (15) we obtain $\int_{\mathcal{X}_{obs}} \delta(x)^2 p(x) dx$, which is not zero, and in general it will be possible to reduce the integral (15) with a different value of $\hat{\theta}$. This best-fitting value depends on the model discrepancy, and will also be different for different $\mathcal{X}_{obs}$ (and even for different densities $p(x)$ of observations over $\mathcal{X}_{obs}$). This is easy to see in the Simple Machine example, Figure 1, where the slope of the best fitting straight line through the origin would clearly be different if we had observations over the range $[4, 6]$ instead of in $[0.2, 4]$.

Consider now the limiting posterior uncertainty around the asymptotic best fitting $\hat{\theta}$. We suppose first that θ is identifiable (using observations over χ_{obs}) in the nonlinear regression model (14). Then the posterior variance will go to zero, so as the sample size increases, the posterior density converges more and more tightly around the best-fitting value and we become more and more certain that θ lies in some arbitrarily small neighbourhood of this estimate. Prior information about θ has no effect because it is ultimately overwhelmed by the evidence from the observations. If θ consists entirely of tuning parameters, then these findings give no cause for concern; calibration will be correctly identifying the best fitting values of these parameters. If, however, there are physical parameters then the posterior inference will in general be biased and unable to identify their true physical values.

Prediction (both interpolation and extrapolation) is equally bad. With an infinitely large sample, the posterior distribution of $\zeta(x)$ is centred on $\eta(x, \hat{\theta})$, with zero variance. Again, as the sample size increases we become increasingly certain that $\zeta(x)$ lies in an arbitrarily small neighbourhood of $\eta(x, \hat{\theta})$, which for almost all x will not equal the true value $\zeta(x)$.

Now consider the case where θ is not identifiable in the model (14). This situation is not uncommon in the use of simulation models, and is sometimes referred to as equifinality; see, for example, Beven and Freer (2001). In this case there will not be a unique $\hat{\theta}$ minimising (15) but a set of such values, and the posterior distribution will in the limit assign probability one to this set. But because of model discrepancy, the true physical value of θ will in general not lie in this set. Consequently, the conclusion in this case is essentially the same: as the sample size increases we become more and more certain that θ lies in an arbitrarily small region around a set of values that does not contain the true value. Prediction is in general unaffected by equifinality, so the conclusion here is exactly as before.

In summary, calibration without explicit recognition of model discrepancy treats all calibration parameters as tuning parameters. In that case it is not possible to learn the true values of physical parameters, nor to predict accurately the real system $\zeta(x)$. With increasing amounts of data we converge with increasing certainty on false values.

4.2 Calibration with model discrepancy

Now suppose that we use the correct model (3). The model discrepancy $\delta(\cdot)$ is unknown and in principle could be almost any function. We will therefore suppose that a prior distribution is assigned for $\delta(\cdot)$ with support that is dense over the space of possible functions. The Gaussian process is one of many forms of prior distribution with this property. In conventional statistics terms, (3) is a nonparametric regression model.

Again consider calibration with an infinite number of observations spread over the subspace $\mathcal{X}_{obs}$. Since the model is nonparametric and can adapt to any possible form of $\delta(\cdot)$ we will learn the true system function $\zeta(x)$ for all $x \in \mathcal{X}_{obs}$. There is therefore no difficulty with interpolation; the posterior distribution of $\zeta(x_0)$ converges to the true value with increasing certainty for all $x_0 \in \mathcal{X}_{obs}$. However, this does not necessarily mean that we can learn the true values of physical parameters or extrapolate reliably.

We can rewrite (2) as

$$\delta(x) = \zeta(x) - \eta(x, \theta) \ . \tag{16}$$

Although we may learn reality $\zeta(\cdot)$ perfectly, for any value of θ there is a $\delta(\cdot)$ given by (16) that agrees perfectly with that reality. There is redundancy in the model; $\zeta(\cdot)$ is identifiable from observations, but θ and $\delta(\cdot)$ are not. Calibration produces a joint posterior distribution for θ and $\delta(\cdot)$ but as the sample size increases we become more and more certain that θ and $\delta(\cdot)$ lie in the manifold

$$\mathcal{M}_\zeta = \{(\theta, \delta(\cdot)) : \eta(x, \theta) + \delta(x) = \zeta(x);\ x \in \mathcal{X}_{obs}\} \tag{17}$$

defined by the true physical system function $\zeta(\cdot)$ and equation (16). All values of θ can be found on that manifold and so the marginal posterior distribution of θ is not degenerate. Even with an infinite quantity of observational data we cannot learn θ precisely.

Figure 7 shows the confounding between θ and $\delta(\cdot)$ in the Simple Machine example for the case of the unconstrained prior with $n = 61$. Figure 7(c) shows the posterior distribution for $\zeta(\cdot)$, obtained by plotting a subsample of the MCMC sample realisations from this distribution. We see that after 61 observations we have effectively learned $\zeta(\cdot)$ over the range of the data with very little posterior uncertainty. Increasing n would simply make the distribution even tighter and eventually degenerate. Figure 7(b) shows the posterior distribution of $\delta(\cdot)$, obtained in the same way by plotting MCMC samples. We see that even with 61 observations we have by no means learnt the true model discrepancy function, and increasing n would not change this plot appreciably. Even with an infinite amount of data we cannot learn $\delta(\cdot)$ because it is not identifiable. Figure 7(a) plots the posterior

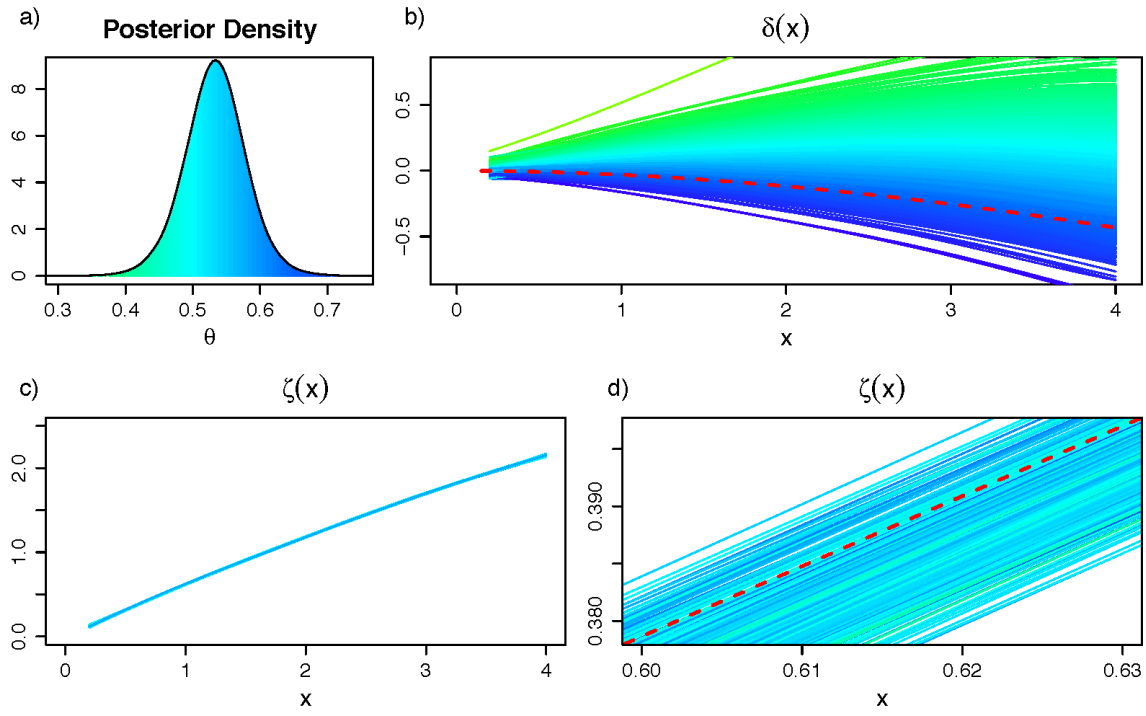

Figure 7: Confounding between θ and $\delta(x)$ for the unconstrained prior and $n = 61$. a) Posterior density of θ showing the color coding of samples from light green (low θ values) to dark blue (high θ values). b) Posterior ensembles of the model-discrepancy $\delta(x)$ color coded by the value of θ. Red dashed line shows the true $\delta(x)$. c) Posterior ensembles of the true process $\zeta(x)$ color coded by the value of θ. d) Close-up of c), red dashed line shows the true $\zeta(x)$.

density of θ and again this will not become appreciably narrower with more data. To show the confounding between θ and $\delta(\cdot)$, the MCMC sample realisations of $\delta(\cdot)$ in panel (b) are coloured with green when they are accompanied by low sampled values of θ, shading to blue where they are accompanied by high values of θ. The colour-coding is shown by shading the density of θ in panel (a). We see a clear colour gradient in panel (b), with high trajectories for $\delta(\cdot)$ corresponding to low values of θ, and vice versa. This is the manifold $\mathcal{M}_\zeta$ for the Simple Machine; when θ is lower, the simulator predictions θx are low and in order for the predicted $\zeta(x)$ to agree with the observations (16) says that $\delta(x)$ must be higher. Figure 7(d) shows part of the posterior distribution shown in panel (c), magnified so that we can see the realisations, which have also been colour-coded in the same way. We do not see a clear colour gradient in this distribution, showing that accurate interpolation of $\zeta(\cdot)$ is obtained for all θ values across the manifold.

Returning to the general case, all points on $\mathcal{M}_\zeta$ have equal likelihood and the posterior distribution is determined here completely by the prior distribution. The posterior density will be high for the true values of θ and $\delta(\cdot)$ only if those values have high prior density relative to other solutions of (16). With a finite sample size, the posterior distribution is not degenerate on $\mathcal{M}_\zeta$ but concentrates increasingly around this set at n increases. We now distinguish between two types of prior distributions.

Case 1. Weak prior information about θ and a diffuse, zero-mean prior distribution for $\delta(\cdot)$.

This was the approach used in Kennedy and O'Hagan (2001). In this case the posterior distribution will be rather flat over the set of $(\theta, \delta(\cdot))$ pairs that give functions $\zeta(\cdot)$ which fit the observations, but will generally assign more probability to pairs where $\delta(x)$ is nearer zero for $x \in \mathcal{X}_{obs}$ since the prior mean of $\delta(x)$ is zero. As a result, the posterior distribution of θ will centre around a 'posterior best fit' value that is similar to $\hat{\theta}$. The posterior variance of θ will not tend to zero but may not be large enough for credible intervals to cover the true physical θ value. A similar situation holds for extrapolation. For x sufficiently far from $\mathcal{X}_{obs}$ the posterior distribution of $\delta(x)$ will be almost the same as the prior distribution, and in particular will have zero mean. Therefore extrapolation will be centred on $\eta(x, \theta)$, with θ set at its posterior best fit value, and so will again be biased. The posterior variance of $\zeta(x)$ will not tend to zero, but again may not be large enough for credible intervals to cover the true value.

The unconstrained prior in the Simple Machine example corresponds to Case 1 and we see these theoretical findings confirmed in the example. In particular, the posterior mean of θ is 0.533, relatively close to the least squares fit value 0.565 from the analysis without

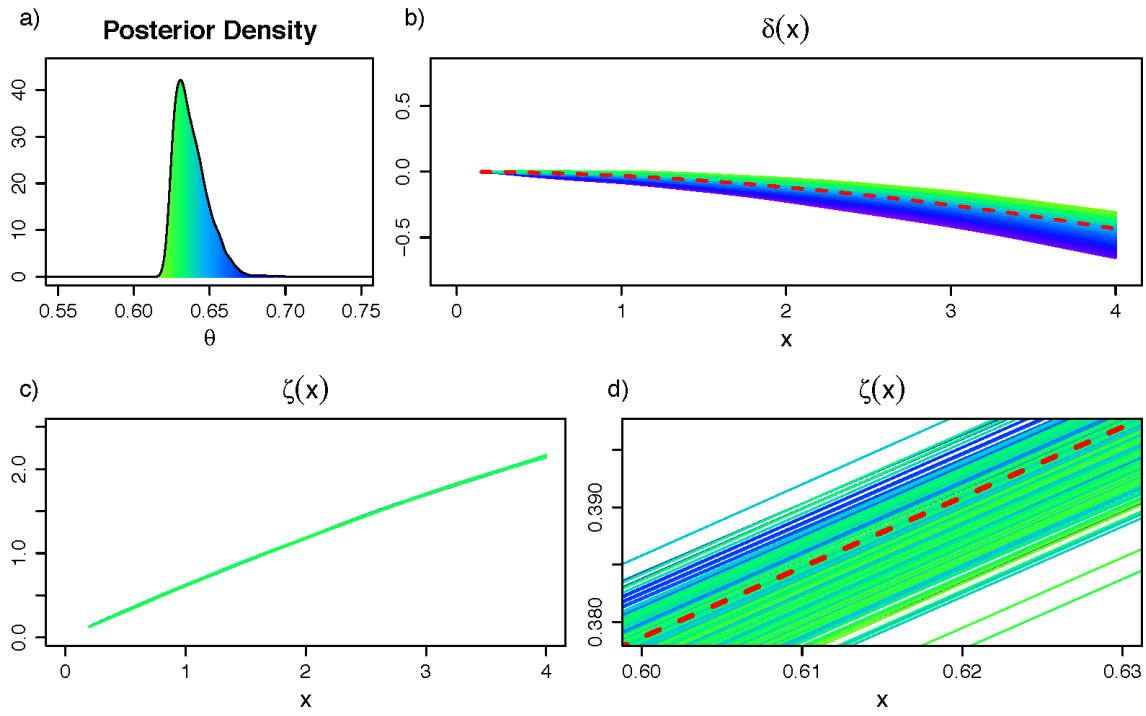

Figure 8: Confounding between θ and $\delta(x)$ for the constrained prior and $n = 61$. a) Posterior density of θ showing the color coding of samples from light green (low θ values) to dark blue (high θ values). b) Posterior ensembles of the model-discrepancy $\delta(x)$ color coded by the value of θ. Red dashed line shows the true $\delta(x)$. c) Posterior ensembles of the true process $\zeta(x)$ color coded by the value of θ. d) Close-up of c), red dashed line shows the true $\zeta(x)$.

model discrepancy and in this instance is even further from the true physical value of 0.65. The posterior distribution of θ does not become degenerate with increasing data, but nevertheless in this example the posterior credible interval does not come close to including the true value. The nonparametric prior distribution allows accurate inference about $\zeta(x)$ within the range of the data (interpolation) but extrapolation is again poor.

Case 2. Realistic prior information about $\delta(\cdot)$ and/or θ.

Stronger prior information will in general lead to a tighter posterior distribution over $\mathcal{M}_\zeta$. Of course, if that information is wrong in the sense that it assigns low prior probability to the true $(\theta, \delta(\cdot))$ pair relative to other points on $\mathcal{M}_\zeta$, then that posterior distribution may again fail to cover the true values. The more accurate the prior information, in the sense of assigning relatively high prior probability to the true $(\theta, \delta(\cdot))$ pair, the closer posterior estimates will be to those true values and the greater the chance of posterior credible intervals covering the true values. Notice that this does not mean that in order to get good posterior inference we have almost to know the true values *a priori*. The data will reduce the space of plausible $(\theta, \delta(\cdot))$ pairs, ultimately to $\mathcal{M}_\zeta$ given enough data, and we only require the prior information to favour the true values over other pairs in that plausible space. However, although we can hope to learn about the true physical parameter θ, accurate extrapolation is more challenging. The posterior distribution of $\zeta(x_0)$ for x_0 outside the observed data now depends on the prior distribution of $\delta(x_0)$ conditional on what we have learnt about $\delta(x)$ in the range of the data.

In the Simple Machine example, the constrained prior is intended to provide the kind of realistic prior information which might be available in a real example, and so corresponds to Case 2. Figure 8 shows the same information for this case as Figure 7 does for the unconstrained prior. Comparing the two figures, we see that panels (c) and (d) are essentially the same in both cases, showing that we learn about $\zeta(\cdot)$ over the range of the data for any nonparametric model discrepancy prior, and that there is no posterior correlation with θ. In panels (a) and (b) the shading again illustrates the strong posterior correlation between θ and $\delta(\cdot)$, but now the posterior distribution of θ in Figure 8(a) is quite different from that in Figure 7(a). It is centred close to the true physical value of 0.65 and comfortably includes that value. Figure 8(b) provides the explanation because we see here the effect of the prior constraints. The posterior distribution of $\delta(\cdot)$ is now quite tightly concentrated on functions that have the right properties and covers the true model discrepancy function (shown in red) over this range.

However, even with this constrained prior, extrapolation to $x_0 = 6$ does not capture the true value of $\zeta(6)$, as we saw in Figure 5. Figure 9 shows what is happening in this example.

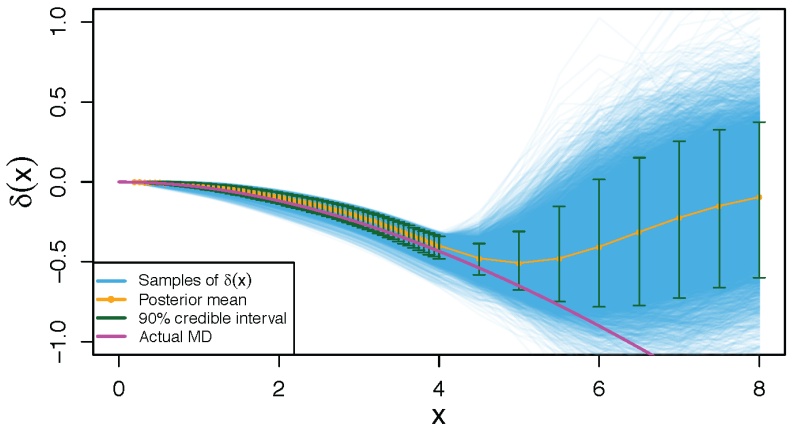

Figure 9: Posterior ensembles of the model-discrepancy, $\delta(x)$, at both observation points and prediction points for the constrained case with $n = 61$.

As soon as x_0 is sufficiently far from 4 (relative to the correlation hyperparameter ψ) so that the $\delta(x_0)$ is independent of $\delta(4)$, the posterior distribution of $\delta(x_0)$ will be the same as its prior distribution, which is centred on zero in this case. So although we are quite accurately capturing the simulator output term θx_0 for the true physical value of θ the observations and the prior distribution have not enabled us to learn about the model discrepancy at $x_0 = 6$.

In order to obtain realistic extrapolation we need realistic prior information about $\delta(\cdot)$, both in the range of the data and out to the control variable values that we wish to predict.

5 Discussion

We have shown, through a simple numerical example and from general theoretical arguments, that calibration without any explicit recognition of model discrepancy is dangerous and leads to poor posterior inference about calibration parameters and poor predictive inference. However, simply introducing model discrepancy with weak prior information, as was done in Kennedy and O'Hagan (2001), is only a partial remedy. It can give accurate interpolation but does not enable learning about physical parameters or extrapolation. Wherever there is interest in learning about the values of physical parameters, or in extrapolating to contexts outside those for which we have observations of the physical system, it is important to incorporate the best and most realistic prior information available about the parameters and the model discrepancy function.

We model the model discrepancy as an additive correlated error term in this paper (a natural approach for the Simple Machine). Other approaches are of course possible, in fact Kennedy and O'Hagan (2001) included a multiplicative model error term as well although that approach has not been widely adopted in practice. Furthermore, the Gaussian Process is not the only way to model a model discrepancy function, although in our view it is one of the more flexible and convenient tools.

In our Simple Machine example, the constrained prior introduced realistic prior information about the nature of the model discrepancy. Without giving any specific prior information about the magnitude of model discrepancy (which might in practice be available), we simply constrained the shape of $\delta(\cdot)$ to be convex and decreasing from $\delta(0) = 0$. The prior information reflected the physical definition of the parameter θ as the gradient of the true process at the origin. We found that constraining the model discrepancy function in this way enabled realistic inference about θ and about $\delta(x)$ for x in the range of the data, i.e. $x \in [0.2, 4]$. However, the prior information was not enough to enable good extrapolation just a short distance from those data, at $x = 6$. We could have added further constraints of the form $\delta'(x) < 0$ for larger x values, but although this would ensure that the posterior mean of $\delta(6)$ did not return to zero it would stay close to $\delta(4)$ and so would have limited value for extrapolation in practice. Alternative Gaussian process models for the model discrepancy might be considered that would more naturally express a prior belief that the magnitude of the model discrepancy should increase with x. For instance, the posterior mean of $\delta(\cdot)$ would not return zero as we extrapolate away from $x = 4$ if the prior discrepancy model were a Gaussian random walk process, and the increasing variance of extroplation in this case would also be more realistic. It is tempting to introduce a drift into such a model with a negative mean for the increments, to be estimated from the data, but this would be confounded with θ and would compromise the latter's physical meaning. A model such as the localised regression model of O'Hagan (1978) could also be considered. The Simple Machine example serves not only to illustrate the dangers of failing to acknowledge model discrepancy but also to show that realistic modelling of model discrepancy is a challenge.

The Simple Machine example, albeit a very simple simulator, does a good job of bringing out the issues we wished to convey in this paper, without adding unnecessary detail. In more complicated calibration/inverse problems the same issues of confounding between model discrepancy and parameters will arise. In addition, more complicated simulators bring even more challenges both computational and by introducing other sources of uncertainty (e.g. emulator error and discretization error) that need to be addressed in concert with model discrepancy and parameter uncertainty. How well these other sources of uncertainty are distinguishable from model discrepancy is not entirely clear and is subject to further research. For example, discretization/numeric error may not be formally distinguishable from model discrepancy. However, this source of error is often well understood and such

prior information may help to separate it from model discrepancy. Currently, discretization error is rarely characterized in a probabilistic way; a notable exception is Chkrebtii et al. (2014).

Prior information about physical parameters can be included, and there are formal techniques of elicitation which can be deployed to express the knowledge of scientists or other experts as probability distributions. However, such information will generally add little because the essence of calibration is generally to learn about parameters whose values are poorly known *a priori*. We believe that formulating prior information about model discrepancy is the key to realistic calibration. In most cases modellers have some idea about what physics or processes the simulator is missing, and we need to use the best judgements of modellers and model users about how the simulator's deficiencies will translate into model discrepancy. However, this is a much more difficult task than eliciting prior distributions for physical parameters. We have illustrated one approach through conditioning a simple Gaussian process, but we have also discussed alternative models for the Simple Machine example which might be able to model the prior knowledge more realistically, and in particular might enable better extrapolation. Prior information can take many forms and other prior modelling approaches will undoubtedly be needed for other contexts. Research is needed into tools for formulating prior knowledge about model discrepancy, and in view of the importance of calibrating simulators we argue that the development of such tools should be a priority area for research in this field.

It must also be recognised that however well we tackle the calibration problem we cannot ever learn the true values of physical parameters, or predict the true process in regions outside $\mathcal{X}_{obs}$, where we have no observational data, to arbitrary accuracy no matter how many observations we make for $x \in \mathcal{X}_{obs}$. The calibration parameters and the model discrepancy function are not identifiable and the accuracy of such inferences is limited by the strength and accuracy of prior information over the manifold $\mathcal{M}_\zeta$ defined by (17). This is a fundamental limitation to the use of mechanistic models.

When prior knowledge is available about what the simulator is missing, one might argue that such information should then be incorporated in the simulator rather than using it to inform a prior distribution on model discrepancy. While changing the simulator may be the right thing to do in many cases, there are also many situation where that is not possible or even advisable. Firstly, it may not be practical to further complicate simulators that take days or weeks to evaluate for one input setting. Secondly, prior knowledge is sometimes of a more qualitative nature which is not easy to model. For example, the prior knowledge that our simulator for the Simple Machine "does not account for losses" is not readily converted into a mathematical form. Thirdly, the available prior knowledge may be uncertain or conflicting (e.g. from different experts). Then rather than putting all our eggs

in one basket by choosing one of the possibilities we take this uncertainty into account in our prior distribution for model discrepancy. We emphasize that we are certainly not advocating ignoring opportunities to improve simulators. On the contrary, modelling is fundamentally an iterative process. By challenging models with observational data we learn about their deficienccies and this information may help the modeller to identify the next improvements to the simulator. It is our view that recognising the existence of model discrepancy, and carefully characterising prior information about the nature of the discrepancy, is not only essential for scientific learning about physical parameters and for extrapolation, it is also an intrinsic part of this iteration.

Here the famous quote of George E. Box comes to mind: "Essentially, all models are wrong, but some are useful" (Box and Draper, 1987). But a model that is wrong can only be useful if we acknowledge the fact that it is wrong. We argue that model discrepancy is an important part of uncertainty quantification and must not be ignored, even though it may be hard to account for.

Acknowledgements

This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

A Analysis without Model discrepancy

For the linear model in (7) and the prior $p(\theta, \sigma_\epsilon^2) \propto \sigma_\epsilon^{-2}$, the posterior distribution of θ is related to the Student's t_{n-1} distribution in the following way:

$$\frac{(X^\intercal X)^{1/2}(\theta - \hat{\theta})}{S} \sim t_{n-1} \tag{18}$$

where $\hat{\theta} = (X^\intercal X)^{-1} X^\intercal \mathbf{Z}$ and $S^2 = \frac{1}{n-1}(\mathbf{Z} - X\hat{\theta})^\intercal(\mathbf{Z} - X\hat{\theta})$.

B MCMC methods for analysis with Model Discrepancy

For the analysis that accounts for model-discrepancy (MD) we obtain approximate samples from the posterior via a Gibbs sampler with a Metropolis-Hastings step for the correlation length ψ. Here we give the full conditional distributions for both the unconstrained approach (Appendix B.1) and the constrained approach (Appendix B.2). First we establish some common notation.

Notation:

$\mathbf{Z}$: Observations. n dim. vector, $n = 11$, 31 or 61.

$\mathbf{x}$: Locations of the observations. n dim. vector, $n = 11$, 31 or 61.

$\boldsymbol{\delta}_o$: MD at points, $\mathbf{x}_o = (0.2, 0.96, 1.72, 2.48, 3.24, 4.00)^\intercal$. $n_o = 6$ dim. vector.

$\boldsymbol{\delta}_r$: MD at the remaining locations, $\mathbf{x}_r$. $n_r = n - n_o$ dim. vector.

$\boldsymbol{\delta}_p$: MD at additional prediction points. n_p dim. vector.

$\boldsymbol{\delta}_q$ $= (\delta'(x_1), \delta'(x_2))^\intercal$: Constrained case only. $x_1 = 0.5$, $x_2 = 1.5$
Derivatives of MD, for inequality constraints. $n_q = 2$ dim. vector.

$\boldsymbol{\delta}_c$ $= (\delta(0), \delta'(0))^\intercal$: Constrained case only.
MD and its derivatives, for equality constraints. $n_c = 2$ dim. vector

We sample the posterior distributions of $\boldsymbol{\delta}_o$ and $\boldsymbol{\delta}_q$ (in the constrained case) and the $\boldsymbol{\delta}_p$ is sampled simultaneously for convenience. The $\boldsymbol{\delta}_c$ vector is not sampled and the $\boldsymbol{\delta}_r$ vector is set equal to the conditional mean given $\boldsymbol{\delta}_o$.

The joint prior distribution of $\boldsymbol{\delta}_o$, $\boldsymbol{\delta}_r$, $\boldsymbol{\delta}_p$, $\boldsymbol{\delta}_q$ and $\boldsymbol{\delta}_c$ is $N\left(\mathbf{0}, \sigma^2\Lambda(\psi)\right)$ where $\Lambda(\psi)$ is calculated using the Gaussian correlation function and ψ represents correlation length. Partitions of $\Lambda(\psi)$ and subsequent conditional matrices will be denoted with o, r, p, q, c as needed. Let $\boldsymbol{\theta} = (\sigma^2, \psi, \sigma_\epsilon^2)$. The notation "$\left[X\right]$" stands for "the probability density function of X".

B.1 Unconstrained GP prior

A complete formulation of the model is the following

$$
\begin{aligned}
\mathbf{Z}|\boldsymbol{\delta}_o, \boldsymbol{\theta} &\sim \mathrm{N}(X\theta + H(\psi)\boldsymbol{\delta}_o, \sigma_\epsilon^2 I) \\
\boldsymbol{\delta}_o|\boldsymbol{\theta} &\sim \mathrm{N}\left(\mathbf{0}, \sigma^2\Lambda(\psi)_{o,o}\right) \\
\left[\theta\right] &\propto 1, \quad \sigma_\epsilon^2 \sim \mathrm{IG}(a_\epsilon, b_\epsilon), \quad \sigma^2 \sim \mathrm{IG}(a, b), \\
\psi &\sim \mathrm{trGamma}_{(0,4)}(a_\psi, b_\psi)
\end{aligned} \tag{19}
$$

where $\text{trGamma}_{(0,4)}$ denotes the Gamma distribution truncated to the interval $(0, 4)$. The matrix $H(\psi)$ is the identity matrix for the rows corresponding to $\mathbf{x}_o$ and equal to $\Lambda(\psi)_{r,o}\Lambda(\psi)_{o,o}^{-1}$ for the other rows, i.e. the conditional mean of $\boldsymbol{\delta}_r$ given $\boldsymbol{\delta}_o$.

Note that for high values of ψ the covariance matrix $\Lambda(\psi)$ becomes numerically singular. Also, θ and $\boldsymbol{\delta}_o$ mix very slowly which is not surprising since there is an non-identifiability issue between these parameters. To deal with this we jointly sample θ, $\boldsymbol{\delta}_o$ and $\boldsymbol{\delta}_p$.

The joint full conditional distribution of $\theta, \boldsymbol{\delta}_o$ and $\boldsymbol{\delta}_p$ given $\mathbf{Z}$ and $\boldsymbol{\theta}$ is $N(\widetilde{\Sigma}^{-1}\mathbf{m}, \widetilde{\Sigma}^{-1})$ where

$$\mathbf{m} = \begin{pmatrix} \frac{1}{\sigma_\epsilon^2} X^\intercal \mathbf{Z} \\ \frac{1}{\sigma_\epsilon^2} H^\intercal \mathbf{Z} \\ \mathbf{0}_{n_p} \end{pmatrix} \quad \text{and} \quad \widetilde{\Sigma} = \begin{pmatrix} \frac{1}{\sigma_\epsilon^2} X^\intercal X & \frac{1}{\sigma_\epsilon^2} X^\intercal A \\ \frac{1}{\sigma_\epsilon^2} A^\intercal X & \frac{1}{\sigma_\epsilon^2} A^\intercal A + \frac{1}{\sigma^2} \Lambda_{op,op}^{-1} \end{pmatrix} \tag{20}$$

where $A = \begin{pmatrix} H & 0_{n \times (n_o + n_p)} \end{pmatrix}$. Note that both H and Λ depend on ψ but that is omitted from the notation here for clarity. By utilizing formulas for the inverse of a partitioned matrixs and the Sherman-Morrison-Woodbury formula (twice) we only need inverses of matrices of size $n_o \times n_o = 6 \times 6$ to calculate $\widetilde{\Sigma}^{-1}$.

The full conditional distributions of the variances σ^2 and σ_ϵ^2 are

$$\sigma^2 | \mathbf{Z}, \boldsymbol{\delta}_o, \theta, \sigma_\epsilon^2, \psi \sim \text{IG}\left(a + \frac{n_o}{2},\ b + \frac{1}{2} ||\boldsymbol{\delta}_o||^2_{\Lambda(\psi)_{o,o}^{-1}} \right) \tag{21}$$

$$\sigma_\epsilon^2 | \mathbf{Z}, \boldsymbol{\delta}_o, \theta, \sigma^2, \psi \sim \text{IG}\left(a_\epsilon + \frac{n}{2},\ b_\epsilon + \frac{1}{2} ||\mathbf{Z} - X\theta - H(\psi)\boldsymbol{\delta_o}||^2 \right) \tag{22}$$

We need a Metropolis-Hastings step for ψ. The full conditional distribution of ψ is

$$\begin{aligned} \left[\psi | \mathbf{Z}, \boldsymbol{\delta}_o, \theta, \sigma^2, \sigma_\epsilon^2\right] \propto & \frac{1}{|\Lambda(\psi)_{o,o}|^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} ||\boldsymbol{\delta}_o||^2_{\Lambda(\psi)_{o,o}^{-1}} \right\} \\ & \times \exp\left\{ -\frac{1}{2\sigma_\epsilon^2} ||\mathbf{Z} - X\theta - H(\psi)\boldsymbol{\delta}_o||^2 \right\} \psi^{a_\psi - 1} e^{-b_\psi \psi} \end{aligned}$$

Let $\omega = \log(\psi)$. Let ω^{t-1} be the last sample for ω and let ω^* be the new proposed value, sampled from $N(\omega^{t-1}, \tau^2)$ truncated above by $\log(4)$, where τ^2 is set in advance. We accept the proposed value with probability $\min(1, r)$ where

$$\begin{aligned} \log(r) = & \frac{1}{2} \log\left(\left| \Lambda(e^{\omega^{t-1}})_{o,o} \right| \right) - \frac{1}{2} \log\left(\left| \Lambda(e^{\omega^*})_{o,o} \right| \right) + (\omega^* - \omega^{t-1}) a_\psi - (e^{\omega^*} - e^{\omega^{t-1}}) b_\psi \\ & - \frac{1}{2\sigma^2} \boldsymbol{\delta}_o^\intercal \left(\Lambda(e^{\omega^*})_{o,o}^{-1} - \Lambda(e^{\omega^{t-1}})_{o,o}^{-1} \right) \boldsymbol{\delta}_o \\ & - \frac{1}{2\sigma_\epsilon^2} \left|\left| \mathbf{Z} - X\theta - H(e^{\omega^*})\boldsymbol{\delta}_o \right|\right|^2 + \frac{1}{2\sigma_\epsilon^2} \left|\left| \mathbf{Z} - X\theta - H(e^{\omega^{t-1}})\boldsymbol{\delta}_o \right|\right|^2 . \end{aligned} \tag{23}$$

B.2 Constrained GP prior

We used a similar Gibbs sampler for the case a constrained GP prior on the model discrepancy, except we condition on $\boldsymbol{\delta}_c = \mathbf{0}$ up front and need to sample the $\boldsymbol{\delta}_q$ vector (the vector of constrained model discrepancy). The prior conditional distribution of $\boldsymbol{\delta}_o$, $\boldsymbol{\delta}_r$, $\boldsymbol{\delta}_p$, $\boldsymbol{\delta}_q$ given $\boldsymbol{\delta}_c = \mathbf{0}$ is $N\left(\mathbf{0}, \sigma^2\Sigma(\psi)\right)$ where

$$\Sigma(\psi) = \Lambda(\psi)_{orpq,orpq} - \Lambda(\psi)_{orpq,c}\Lambda(\psi)_{cc}^{-1}\Lambda(\psi)_{c,orpq} \ .$$

The $\Lambda(\psi)$ covariance matrix is calculated by using the Gaussian correlation function $c(x_1, x_2)$ in (10) and the covariance between derivatives of $\delta(x)$ when appropriate:

$$\mathrm{Cov}(\delta^1(x_1), \delta(x_2)) = -\sigma^2 \frac{2(x_1 - x_2)}{\psi^2} \exp\left\{-\frac{(x_1 - x_2)^2}{\psi^2}\right\} \tag{24}$$

$$\mathrm{Cov}(\delta^1(x_1), \delta^1(x_2)) = \sigma^2 \frac{2}{\psi^2} \exp\left\{-\frac{(x_1 - x_2)^2}{\psi^2}\right\} \left(1 - \frac{2(x_1 - x_2)^2}{\psi^2}\right) \tag{25}$$

A complete formulation of the model is the following:

$$\begin{aligned}
\mathbf{Z}|\boldsymbol{\delta}_o, \boldsymbol{\delta}_c = \mathbf{0}, \boldsymbol{\theta} &\sim \mathrm{N}(X\theta + H(\psi)\boldsymbol{\delta}_o, \sigma_\epsilon^2 I) \\
\boldsymbol{\delta}_o|\boldsymbol{\delta}_q, \boldsymbol{\delta}_c = \mathbf{0}, \boldsymbol{\theta} &\sim \mathrm{N}\left(\Sigma(\psi)_{oq}\Sigma(\psi)_{qq}^{-1}\boldsymbol{\delta}_q, \sigma^2(\Sigma(\psi)_{oo} - \Sigma(\psi)_{oq}\Sigma(\psi)_{qq}^{-1}\Sigma(\psi)_{qo})\right) \\
\boldsymbol{\delta}_q|\boldsymbol{\delta}_c = \mathbf{0}, \boldsymbol{\theta} &\sim \mathrm{trN}_{(-\infty,0)}(\mathbf{0}, \sigma^2\Sigma(\psi)_{qq}) \\
\left[\theta\right] \propto 1, \quad \sigma_\epsilon^2 &\sim \mathrm{IG}(a_\epsilon, b_\epsilon), \quad \sigma^2 \sim \mathrm{IG}(a, b), \\
\psi &\sim \mathrm{trGamma}_{(0,4)}(a_\psi, b_\psi)
\end{aligned} \tag{26}$$

The matrix $H(\psi)$ is the identity matrix for the rows corresponding to $\mathbf{x}_o$ and equal to $\Sigma(\psi)_{r,o}\Sigma(\psi)_{o,o}^{-1}$ for the other rows, i.e. $\boldsymbol{\delta}_r$ is set equal to its conditional mean given $\boldsymbol{\delta}_o$.

Unlike in the unconstrained case we update the $\boldsymbol{\delta}$ vectors and θ separately. The full conditional distribution $\left[\boldsymbol{\delta}_o, \boldsymbol{\delta}_p, \boldsymbol{\delta}_q|\mathbf{Z}, \boldsymbol{\delta}_c = \mathbf{0}, \boldsymbol{\theta}\right]$ is normal with covariance matrix and mean

$$\tilde{\Sigma} = \sigma^2\Sigma - \sigma^4\Sigma \begin{pmatrix} H^\intercal(\sigma_\epsilon^2 I + \sigma^2 H\Sigma_{oo}H^\intercal)^{-1}H & 0_{n_o \times (n_p+n_q)} \\ 0_{(n_p+n_q)\times n_o} & 0_{(n_p+n_q)\times(n_p+n_q)} \end{pmatrix} \Sigma \tag{27}$$

$$\tilde{\boldsymbol{\mu}} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} \tilde{\Sigma}_{oo} \\ \tilde{\Sigma}_{po} \\ \tilde{\Sigma}_{qo} \end{pmatrix} H^\intercal(\mathbf{Z} - X\theta) \ . \tag{28}$$

where Σ_{oo} is the rows and columns of Σ that correspond to $\boldsymbol{\delta}_o$, i.e. the covariance matrix of $\boldsymbol{\delta}_o|\boldsymbol{\delta}_c = \mathbf{0}, \boldsymbol{\theta}$. The dependence of H and Σ on ψ has been omitted from the notation. Again, by using the Sherman-Morrison-Woodbury formula we only need to calculate inverses of

matrices of size 6×6.

We update $\boldsymbol{\delta}_o$, $\boldsymbol{\delta}_p$ and $\boldsymbol{\delta}_q$ in two steps.

Step 1: Sample $\left[\boldsymbol{\delta}_q | \boldsymbol{\delta}_q \in \mathcal{Q}, \mathbf{Z}, \boldsymbol{\delta}_c = \mathbf{0}, \boldsymbol{\theta}\right]$, where $\mathcal{Q} = (-\infty, 0) \times (-\infty, 0)$. We do this one element at a time:

1. Sample $d_2 \sim \text{trN}_{(-\infty,0)}(\tilde{\mu}_{q,2}, \tilde{\Sigma}_{qq,22})$
2. Sample $d_1 \sim \text{trN}_{(-\infty,0)}\left(\tilde{\mu}_{q,1} + \tilde{\Sigma}_{qq,12}/\tilde{\Sigma}_{qq,22}(d_2 - \tilde{\mu}_{q,2}), \tilde{\Sigma}_{qq,11} - \tilde{\Sigma}^2_{qq,12}/\tilde{\Sigma}_{qq,22}\right)$

Operationally, if the mean $\tilde{\mu}_{q,j}$ is negative we sample the normal distribution until we get a negative number, otherwise we use the rejection sampling algorithm of Robert (1995).

Step 2: Sample $\left[\boldsymbol{\delta}_o, \boldsymbol{\delta}_p | \boldsymbol{\delta}_q = \mathbf{d}, \boldsymbol{\delta}_c = \mathbf{c}, \mathbf{Z}, \boldsymbol{\theta}\right]$ where $\mathbf{d} = (d_1, d_2)^\intercal$ from step 1. This distribution is normal with mean and variance

$$\begin{aligned}
\bar{\boldsymbol{\mu}} &= \mathbf{m}_{op} + \tilde{\Sigma}_{op,q}\tilde{\Sigma}^{-1}_{q,q}(\mathbf{d} - \mathbf{m}_q) \\
\bar{\Sigma} &= \tilde{\Sigma}_{op,op} - \tilde{\Sigma}_{op,q}\tilde{\Sigma}^{-1}_{q,q}\tilde{\Sigma}_{q,op}
\end{aligned}$$

Other parameters:
The full conditional distribution of θ is normal with mean and variance

$$\mu_\theta = (X^\intercal X)^{-1} X^\intercal (\mathbf{Z} - H(\psi)\boldsymbol{\delta}_o) \quad \text{and} \quad \sigma^2_\theta = \sigma^2_\epsilon (X^\intercal X)^{-1} \ .$$

The full conditional distributions of the variances σ^2 and σ^2_ϵ are the following:

$$\begin{aligned}
\sigma^2 | \mathbf{Z}, \boldsymbol{\delta}_o, \boldsymbol{\delta}_q, \boldsymbol{\theta}, \sigma^2_\epsilon, \psi &\sim \text{IG}\left(a + \frac{n_o + n_q}{2},\ b + \frac{1}{2}\left|\left|\begin{bmatrix}\boldsymbol{\delta}_o \\ \boldsymbol{\delta}_q\end{bmatrix}\right|\right|^2_{\Sigma^{-1}_{oq,oq}(\psi)}\right) \\
\sigma^2_\epsilon | \mathbf{Z}, \boldsymbol{\delta}_o, \boldsymbol{\delta}_q, \boldsymbol{\theta}, \sigma^2, \psi &\sim \text{IG}\left(a_\epsilon + \frac{n}{2},\ b_\epsilon + \frac{1}{2}||\mathbf{Z} - X\theta - H\boldsymbol{\delta_o}||^2\right) \ .
\end{aligned}$$

We do a Metropolis-Hastings step for ψ and, as before, we use a normal proposal distribution for $\omega = \log(\psi)$. We accept a proposed value ω^* over the last value ω^{t-1} with probability

$\min(1, r)$ where

$$\begin{aligned}\log(r) = &\frac{1}{2}\log\left(\left|\Sigma(e^{\omega^{t-1}})_{oq,oq}\right|\right) - \frac{1}{2}\log\left(\left|\Sigma(e^{\omega^*})_{oq,oq}\right|\right) + (\omega^* - \omega^{t-1})a_\psi - (e^{\omega^*} - e^{\omega^{t-1}})b_\psi \\ &- \frac{1}{2\sigma^2}\begin{bmatrix}\boldsymbol{\delta}_0 \\ \boldsymbol{\delta}_q\end{bmatrix}^\intercal \left(\Sigma(e^{\omega^*})^{-1}_{oq,oq} - \Sigma(e^{\omega^{t-1}})^{-1}_{oq,oq}\right)\begin{bmatrix}\boldsymbol{\delta}_0 \\ \boldsymbol{\delta}_q\end{bmatrix} \\ &- \frac{1}{2\sigma_\epsilon^2}\left|\left|\mathbf{Z} - X\theta - H(e^{\omega^*})\boldsymbol{\delta}_o\right|\right|^2 + \frac{1}{2\sigma_\epsilon^2}\left|\left|\mathbf{Z} - X\theta - H(e^{\omega^{t-1}})\boldsymbol{\delta}_o\right|\right|^2 .\end{aligned}$$

References

Adler, R. J. (2010), *The Geometry of Random Fields*, SIAM.

Apley, D. W., Liu, J., and Chen, W. (2006), "Understanding the Effects of Model Uncertainty in Robust Design With Computer Experiments," *Journal of Mechanical Design*, 128, 945–958.

Arhonditsis, G. B., Papantou, D., Zhang, W., Perhar, G., Massos, E., and Shi, M. (2008), "Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management," *Journal of Marine Systems*, 73, 8–30.

Bayarri, M. J., Berger, J. O., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Lin, C., and Tu, J. (2009), "Predicting Vehicle Crashworthiness: Validation of Computer Models for Functional and Hierarchical Data," *Journal of the American Statistical Association*, 104, 929–943.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., and Tu, J. (2007), "A framework for validation of computer models," *Technometrics*, 49, 138–154.

Beven, K. J. and Freer, J. (2001), "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems," *Journal of Hydrology*, 249, 11–29.

Box, G. E. P. and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, John Wiley & Sons, New York, NY.

Chkrebtii, O. A., Campbell, D. A., Girolami, M. A., and Calderhead, B. (2014), "Bayesian Uncertainty Quantification for Differential Equations," In Review. Available at arXiv:1306.2365v2.

Da Veiga, S. and Marrel, A. (2012), "Gaussian process modeling with inequality constraints," *Annales de la Faculté des Sciences de Toulouse*, 21, 529–555.

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

Gramacy, R. B. and Lee, H. K. H. (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association*, 103, 1119–1130.

Habib, S., Heitmann, K., Higdon, D., Nakhleh, C., and Williams, B. (2007), "Cosmic calibration: Constraints from the matter power spectrum and the cosmic microwave background," *Physical Review D*, 76, 083503.

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

Higdon, D. M., Kennedy, M. C., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004), "Combining field data and computer simulations for calibration and prediction," *SIAM Journal on Scientific Somputing*, 26, 448–466.

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

Lee, H. K. H., Sansó, B., Zhou, W., and Higdon, D. M. (2008), "Inference for a Proton Accelerator Using Convolution Models," *Journal of the American Statistical Association*, 103, 602–613.

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. M. H., and Webb, M. J. (2007), "A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles," *Philos Transact A Math Phys Eng Sci*, 365, 1993–2028.

O'Hagan, A. (1978), "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society B*, 40, 1–42.

— (1992), "Some Bayesian Numerical Analysis," in *Baysian Statistics 4*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 345–363.

Qian, P. Z. G. and Wu, C. F. J. (2008), "Bayesian Hierarchical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments," *Technometrics*, 50, 192–204.

Reichert, P. and Mieleitner, J. (2009), "Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters," *Water Resources research*, 45, W10402, doi:10.1029/2009WR007814.

Riihimäki, J. and Vehtari, A. (2010), "Gaussian processes with monotonicity information," *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 9, 645–652.

Robert, C. P. (1995), "Simulation of truncated normal variables," *Statistics and Computing*, 5, 121–125.

Rougier, J. (2007), "Probabilistic inference for future climate using an ensemble of climate model evaluations," *Climatic Change*, 81, 247–264.

Sansó, B. and Forest, C. (2009), "Statistical calibration of climate system properties," *Journal of the Royal Statistical Society C*, 58, 485–503.

Stainforth, D. A., Allen, M. R., Tredger, E. R., and Smith, L. A. (2007), "Confidence, uncertainty and decision-support relevance in climate predictions," *Philosophical Transactions of The Royal Society A - Mathematical Physical and Engineering Sciences*, 365, 2145–2161.

Strong, M., Oakley, J. E., and Chilcott, J. (2012), "Managing structural uncertainty in health economic decision models: a discrepancy approach," *Journal of the Royal Statistical Society C*, 61, 25–45.

Unal, C., Williams, B., Hemez, F., Atamturktur, S. H., and McClure, P. (2011), "Improved best estimate plus uncertainty methodology, including advanced validation concepts, to license evolving nuclear reactors," *Nuclear Engineering and Design*, 241, 1813–1833.

Wang, X. and Berger, J. O. (2011), "Estimating Shape Constrained Functions Using Gaussian Processes," in *JSM Proceedings, Section on Nonparametric Statistics*, American Statistical Association, Alexandria, VA, pp. 5162–5171.