Learning of model discrepancy for structural dynamics applications using Bayesian history matching

IOP Conf. Series: Journal of Physics: Conf. Series **1264** (2019) 012052 doi:10.1088/1742-6596/1264/1/012052

Learning of model discrepancy for structural dynamics applications using Bayesian history matching

P Gardner, T J Rogers, C Lord and R J Barthorpe

University of Sheffield, Department of Mechanical Engineering, Mappin Street, Sheffield S1 3JD, UK

E-mail: pagardner1@sheffield.ac.uk

Abstract. Calibration of computer models for structural dynamics is often an important task in creating valid predictions that match observational data. However, calibration alone will lead to biased estimates of system parameters when a mechanism for model discrepancy is not included. The definition of model discrepancy is the mismatch between observational data and the model when the 'true' parameters are known. This will occur due to the absence and/or simplification of certain physics in the computer model. Bayesian History Matching (BHM) is a 'likelihood-free' method for obtaining calibrated outputs whilst accounting for model discrepancies, typically via an additional variance term. The approach assesses the input space, using an emulator of the complex computer model, and identifies parameter sets that could have plausibly generated the target outputs. In this paper a more informative methodology is outlined where the functional form of the model discrepancy is inferred, improving predictive performance. The algorithm is applied to a case study for a representative five storey building structure with the objective of calibrating outputs of a finite element (FE) model. The results are discussed with appropriate validation metrics that consider the complete distribution.

1. Introduction

Bayesian History Matching (BHM) is a 'likelihood-free' method for calibrating computer models (here defined as simulators) under the assumption of model discrepancy, i.e. given the simulator was evaluated with the 'true' parameter set there would still be a difference between these predictions and observed data. This is important as without considering this source of uncertainty parameter inferences and output predictions will be biased. The approach is 'likelihood-free' meaning that input and output combinations can be removed and added iteratively without invalidating the analysis. Furthermore, this means that the considered parameter domain can be truncated based on physical understanding of the parameters, reducing non-identifiability issues and non-physical inferences.

The technique, originally developed in the oil industry [1], has been applied to fields such as Galaxy formation [2, 3], complex social models of HIV transfer in populations [4, 5] and climate science [6, 7]. However, the approach has not been investigated for structural dynamics problems to the authors' knowledge. In addition, the methodology does not contain a mechanism for inferring the functional model discrepancy form after calibration. Consequently, this paper proposes an importance sampling based technique for inferring functional model discrepancies,

modelled as Gaussian Processes (GP)s, when a posterior distribution for the parameters is known. The approach is applied to a representative building structure whereby it is shown to be effective in identifying model discrepancy.

2. Bayesian History Matching

BHM seeks to calibrate a statistical model of the form shown in Eq. (1).

$$z_j(\boldsymbol{x}) = \eta_j(\boldsymbol{x}, \boldsymbol{\theta}) + \delta_j + e_j \tag{1}$$

Where $z_j(\boldsymbol{x})$ is the jth observational output given inputs $\boldsymbol{x}$, $\eta_j(\boldsymbol{x}, \boldsymbol{\theta})$ is the jth simulator given $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$. The model discrepancy and observational uncertainty are δ_j and e_j respectively. The model assumes that the simulator, discrepancy and observational uncertainty are independent and does not seek to define the model discrepancy's functional form.

In order to calibrate Eq. (1) the parameter space of the simulator is explored in iterations called waves. During a wave simulator outputs are assessed for different parameter combinations using an implausibility metric and discarded if above a threshold value T. Due to the method requiring assessment of a large parameter space a computationally efficient emulator is utilised reducing the computational burden of evaluating the simulator. Here a GP emulator is implemented; a Bayesian, non-parametric, regression model. A key benefit of using a GP emulator is that it will fit known simulator runs exactly whilst inferring code uncertainty when the emulator predicts away from known simulator runs — in keeping with the brevity of this paper the reader is referred to [8, 9] for more mathematical details on GPs. The GP emulator is constructed as in Eq. (2).

$$\eta_j(\boldsymbol{x}, \boldsymbol{\theta}) \sim \mathcal{GP}_j\left(m(\boldsymbol{x}, \boldsymbol{\theta}), k((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}')\right) \tag{2}$$

Where the emulator GP prior is defined by a mean $m(\boldsymbol{x}, \boldsymbol{\theta})$ and covariance function $k((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}'))$ which have a set of hyperparameters $\boldsymbol{\phi}_\eta$. The predictive GP emulator mean $\mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)$ and code uncertainty, $V_c(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbb{V}\left(\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)$ are then incorporated in an implausibility metric. The formulation stated in Eq. (2) assumes univariate GP emulators for each output, however multivariate GPs could be implemented [10]. The implausibility metric assesses the distance between observations and simulator outputs, weighted by the process's uncertainties, defined in Eq. (3).

$$I_j(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{|z_j(\boldsymbol{x}) - \mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)|}{\left(V_{o,j} + V_{m,j} + V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta})\right)^{1/2}} \tag{3}$$

Where, V_o, V_m and $V_c(\boldsymbol{x}, \boldsymbol{\theta})$ are the variances associated with the observational, model discrepancy and code uncertainties respectively. By including code uncertainty $V_c(\boldsymbol{x}, \boldsymbol{\theta})$ into Eq. (3) parameter space is retained until the emulator predictions are more certain for that particular parameter region. The observational uncertainty V_o can often be estimated from observational data, although expert judgement may also be used. Model discrepancy uncertainty V_m can be more challenging to define, but should be elicited from expert judgement.

The implausibility metric presented in Eq. (3) provides a quantity for every parameter combination, input and output, however a single value is required for each parameter combination in order to decide whether it should be removed. Several extensions of the implausibility metric that deal with multiple outputs and inputs can be considered such as a maximum or multivariate implausibility metric [4]. Here a multivariate implausibility metric for either the inputs or outputs, Eqs. (4) and (5) is utilised. This is equivalent to taking the Mahalanobis distance, standard practice in outlier analysis [11], which assesses the Euclidean distance of the principle components. The maximum can be taken over either

Algorithm 1 Bayesian History Matching for Wave k

$\boldsymbol{\theta}^k \sim$ GMLHC ▷ Draw parameters from GMLHC
$\boldsymbol{y}^k = \eta(\boldsymbol{x}, \boldsymbol{\theta}^k)$ ▷ Run the simulator at parameters
Draw n samples $\boldsymbol{\theta}_s^k \sim \mathcal{U}\left(\min(\boldsymbol{\theta}^k), \max(\boldsymbol{\theta}^k)\right)$ ▷ Sample parameter space
for $j = 1$: no. of outputs **do**
  Train and validate $\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}^k\right)$ ▷ Train and validate emulators
  $\left[\mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}_s^k\right)\right), V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta}_s^k)\right] = \mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}_s^k\right)$ ▷ Predictions at n samples of $\boldsymbol{\theta}^k$
  Calculate $I_j(\boldsymbol{x}, \boldsymbol{\theta}_s^k)$ ▷ Assess implausibility of samples
end for
Calcualate $I_{max}(\boldsymbol{\theta}_s^k)$
for $m = 1 : n$ **do**
  if $I_{max}(\boldsymbol{\theta}_{s,m}^k) < T$ **then**
    $\boldsymbol{\theta}_{nI}^k = \boldsymbol{\theta}_{s,m}^k$ ▷ Keep non-implausible samples
  end if
end for
bounds $= [\min(\boldsymbol{\theta}_{nI}^k), \max(\boldsymbol{\theta}_{nI}^k)]$ ▷ Obtain new GMLHC bounds
if any $(V_{c,j}^k(\boldsymbol{x}, \boldsymbol{\theta}) < (V_{o,j} + V_{m,j}))$ **or** isempty$(\boldsymbol{\theta}_{nI}^k)$ **then**
  Stop ▷ Stop if stopping criteria are met
end if

Eqs. (4) and (5) to collapse the metric to a single value for each parameter combination; where $V_j(\boldsymbol{x}, \boldsymbol{\theta}) = V_{o,j} + V_{m,j} + V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta})$.

$$I_{multi}(\boldsymbol{\theta})_j = \left(z_j(\boldsymbol{x}) - \mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)\right)^{\mathsf{T}} \left(V_j(\boldsymbol{x}, \boldsymbol{\theta})\right)^{-1} \left(z_j(\boldsymbol{x}) - \mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)\right) \tag{4}$$

$$I_{multi}(\boldsymbol{x}, \boldsymbol{\theta}) = \left(z_j(\boldsymbol{x}) - \mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}_*, \boldsymbol{\theta}_*\right)\right)\right)^{\mathsf{T}} \left(V_j(\boldsymbol{x}, \boldsymbol{\theta})\right)^{-1} \left(z_j(\boldsymbol{x}) - \mathbb{E}\left(\mathcal{GP}_j\left(\boldsymbol{x}_*, \boldsymbol{\theta}_*\right)\right)\right) \tag{5}$$

The parameter space is excluded when the implausibility metric is above a threshold T, where the rejection criteria for a particular parameter combination θ is defined in Eq. (6).

$$I(\theta) \begin{cases} \leq T & \text{if } \theta \in \boldsymbol{\theta}_{nI}, \\ > T & \text{if } \theta \in \boldsymbol{\theta}_I \end{cases} \tag{6}$$

The threshold value depends on the type of implausibility metric being considered. For individual implausibilities $I_j(\boldsymbol{x}, \boldsymbol{\theta})$ a sensible threshold T is Pukelsheim's 3σ rule — stating that any continuous unimodal distribution will contain at least 99.5% of probability mass within three standard deviations away from the mean [12]. For multivariate implausibilities the threshold T is set as a high percentile (i.e. $\alpha > 95\%$) from a chi-squared distribution with either j or the input size $\boldsymbol{x}$ degrees of freedom, i.e. $T = F_{\chi^2}^{-1}(\alpha)$ — the output from a chi-squared quantile function. This can be thought of as performing a frequentist hypothesis test on the parameter combination, using a chi-squared (χ^2) test.

The parameter space in each iteration is sampled in order to assess the rejection criteria. Here samples are drawn from a uniform distribution bounded by the initial parameter domain. This works effectively with a Latin Hypercube Design (LHD) based approach. In this scenario the initial parameter space bounds are used, in conjunction with a simulator budget, to construct a LHD — here a Generalised Maximum Latin Hypercube Design (GMLHD) is used [13]. An emulator is constructed from the simulator runs and its output assessed at parameter combinations sampled from a uniform distribution. A set of these sample parameters are rejected

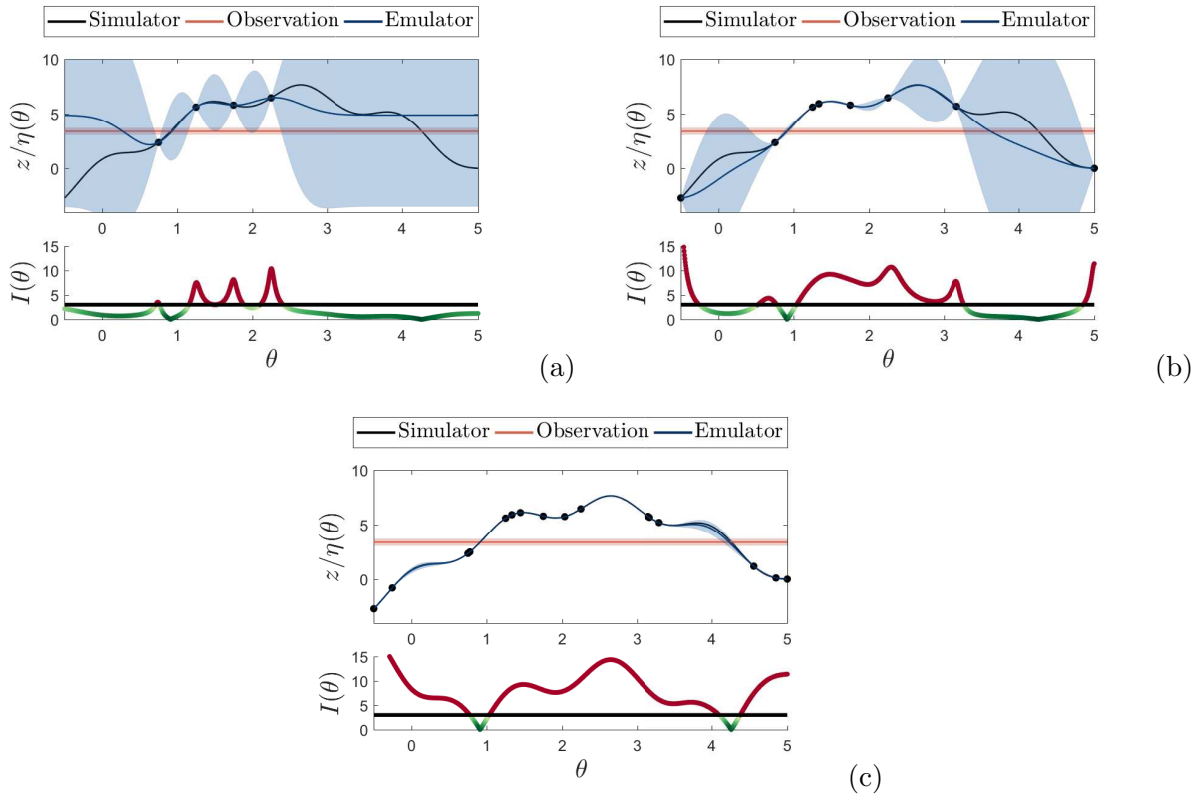


Figure 1. BHM waves $k = 1, 2, 4$ for the numerical example. Top panels show the observational data with $\pm\sqrt{V_o + V_m}$ shaded region against the simulator and emulator predictions, trained using the simulator runs $\eta(\boldsymbol{\theta}^k)$ $(\cdot)$. The bottom panels show the implausibility $I(\boldsymbol{\theta}_s^k)$ against the threshold $T = 3$, where green regions are non-implausible and red implausible. Panel (a), (b) and (c) show waves $k = 1, 2, 4$ respectively.

based on the implausibility metric and criteria, with the bounds of the non-implausible region determined. A new wave can then be run with a LHD constructed from these bounds.

Finally, a stopping criteria is constructed, based on two outcomes; all the space is deemed implausible or the emulator variance in the non-implausible region is less than the remaining uncertainties, i.e. $V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta}_{nI}) < V_{o,j} + V_{m,j}$, which indicates that the emulator is at least as certain about its predictions as the modeller is with the uncertainties due to model discrepancy and observation variability. The stated approach to BHM can be defined in Algorithm 1.

To illustrate BHM, Algorithm 1 is applied to a simple numerical example — see Fig. 1 (where the sampling stage is replaced with a uniform grid). The examples shows that during each wave new simulator evaluations are added, decreasing the code uncertainty of the emulator. This in turn leads to increased parts of the parameter space begin discarded, until the implausibility metric identifies two non-implausibility regions for this example, i.e. the areas where the simulator crosses the observational uncertainty bounds.

2.1. Approximate Posterior Sampling

Approximate posterior samples can be obtained from the non-implausible space identified from the final wave using an importance sampling approach. Importance sampling is a method for obtaining unbiased estimates of expectation integrals, but can also be formulated to approximate a posterior density $p(\boldsymbol{\theta} \,|\, \boldsymbol{z}) = p(\boldsymbol{z} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{z})$ when the evidence $p(\boldsymbol{z})$ cannot be calculated.

The technique samples a set of independent draws from an unnormalised proposal distribution $\boldsymbol{\theta}_q \sim \tilde{q}$ (where $\tilde{q}$ is the unnormalised proposal) which are subsequently used to form a set of unnormalised weights, shown in Eq. (7). The posterior distribution is then estimated by normalising the weights as presented in Eq. (8).

$$\tilde{w}(\boldsymbol{\theta}_q) = \frac{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_q) p(\boldsymbol{\theta}_q)}{\tilde{q}(\boldsymbol{\theta}_q)} \tag{7}$$

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{z}) \approx \frac{\tilde{w}(\boldsymbol{\theta}_q)}{\frac{1}{n}\sum_{i=1}^{n} \tilde{w}(\boldsymbol{\theta}_q)} \tag{8}$$

However, as BHM does not involve a likelihood, an approximation is formed as defined in Eq. (9) — the product of multivariate Gaussian distributions over $z(\boldsymbol{x})$ for the set of inputs $\boldsymbol{x}$. This assumes that the sources of uncertainty are approximately normally distributed. The proposal distribution can be formulated as a multivariate Gaussian distribution as presented in Eq. (10).

$$p(\boldsymbol{z} \,|\, \boldsymbol{\theta}) \approx L(\boldsymbol{\theta}) = \prod_{j=1}^{M} \mathcal{N}\left(\boldsymbol{z}(\boldsymbol{x}) \,|\, \mathbb{E}_j\left(\mathcal{GP}\left(\boldsymbol{x}\,, \boldsymbol{\theta}\right)\right), V_j(\boldsymbol{x}, \boldsymbol{\theta})\right) \tag{9}$$

$$\tilde{q}(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta} \,|\, \mu_{nI}, \kappa\Sigma_{nI}\right) \tag{10}$$

Where μ_{nI} and Σ_{nI} are the sample mean and variance-covariance from the non-implausible set after the last wave, and κ is an inflation parameter to ensure good coverage of the space. The choice of prior $p(\boldsymbol{\theta})$ depends on the modellers beliefs from the last wave. However it is often reasonable to assume a constant prior over the final non-implausible set, as it is often a fraction of the original parameter domain. This means the weights in Eq. (8) become $\tilde{w} = L(\boldsymbol{\theta}_q)/\tilde{q}(\boldsymbol{\theta}_q)$ where $\boldsymbol{\theta}_q$ are a number of samples from $\tilde{q}$, with the constant prior essentially truncating the proposal samples to be within the final non-implausible domain.

Lastly the approximate posterior from Eq. (8) can be re-sampled in order generate direct samples from the posterior. This involves drawing N_q samples where the probability of occurrence is defined by the normalised weights $w(\boldsymbol{\theta}_q) = \tilde{w}(\boldsymbol{\theta}_q)/\frac{1}{n}\sum_{i=1}^{n} \tilde{w}(\boldsymbol{\theta}_q)$.

3. Model Discrepancy Inference via Importance Sampling

BHM provides a mechanism for calibrating additive model discrepancy and identifying the approximate parameter posterior. The method does not provide a mechanism for inferring the model discrepancy functional form and correcting output predictions. Here an importance sampling methodology is proposed in order to infer model discrepancy uncertainty, assumed to be distributed as a GP.

Once the approximate posterior distribution is inferred samples $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \,|\, Z, \boldsymbol{x}_z)$ can be propagated through the emulator such that samples of the calibrated simulator output distribution are obtained $p(\boldsymbol{y}_{*,j}^{(i)} \,|\, \boldsymbol{x}_*, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\phi}}_{\eta,j})$ — $\hat{\boldsymbol{\phi}}_{\eta,j}$ are the jth emulator's Maximum Likelihood Estimate (MLE) estimate of the hyperparameters and Z is a matrix of the outputs $z_j(\boldsymbol{x})|_{j=1:N_{out}}$. Here it is assumed that all the emulators accurately capture the simulators functional form meaning the hyperparameters $\hat{\boldsymbol{\phi}}_{\eta,j}$ can be treated as fixed rather than being marginalised; as already assumed within BHM. For this reason $\hat{\boldsymbol{\phi}}_{\eta,j}$ are dropped from notation.

The desired calibrated and bias corrected distribution for the jth output is $p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z)$. To obtain samples from this distribution GP models that map from $\boldsymbol{y}_j$ to $\boldsymbol{z}_j$ given $\boldsymbol{\theta}$ are constructed such that $p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}, \boldsymbol{\phi}_{\delta,j})$ is formed. The unconditional distribution requires marginalising out the model discrepancy hyperparameters $\boldsymbol{\phi}_{\delta,j}$ and parameters $\boldsymbol{\theta}$ as shown in Eq. (11).

Algorithm 2 Importance Sampling Model Discrepancy Inference

for $j = 1 : N_{out}$ **do**

Training;
for $i = 1 : N_s$ **do**
Predict $p(\boldsymbol{y}_j^{(i)} \,|\, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\phi}}_{\eta,j})$
for $k = 1 : N_\phi$ **do**
Sample $\boldsymbol{\phi}_{\delta,j}^{(i,k)} \sim q\left(\boldsymbol{\phi}\right)_{\delta,j}$
$\tilde{w}_j^{(i,k)} = p(\boldsymbol{z}_j \,|\, \boldsymbol{y}_j^{(i)}, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \boldsymbol{\phi}_{\delta,j}^{(i,k)})$
end for
end for
Normalise weights $w_j^{(i,k)} = \tilde{w}_j^{(i,k)} / \sum_{i=1}^{N_s} \sum_{k=1}^{N_\phi} \tilde{w}_j^{(i,k)}$

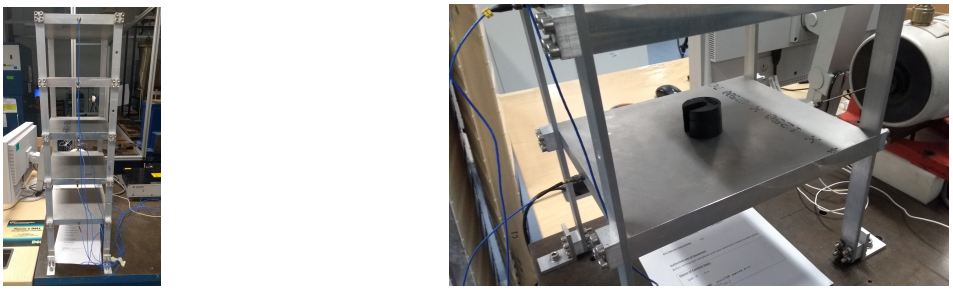
Prediction;
for $i = 1 : N_s$ **do**
Predict $p(\boldsymbol{y}_{*,j}^{(i)} \,|\, \boldsymbol{x}_*, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\phi}}_{\eta,j})$
for $k = 1 : N_\phi$ **do**
Predict $\hat{\boldsymbol{z}}_{*,j}^{(i,k)}$ and $\Sigma_{z*,j}^{(i,k)}$ from $p(\boldsymbol{z}_{*,j}^{(i,k)} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \boldsymbol{\phi}_{\delta,j}^{(i,k)})$
end for
end for
Predict the approximation of $p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z)$ via Eqs. (13) and (14)
end for

$$p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z) = \int \left(\int p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}, \boldsymbol{\phi}_{\delta,j}) p(\boldsymbol{\phi}_{\delta,j}) \, \mathrm{d}\boldsymbol{\phi}_{\boldsymbol{\delta}} \right) \, p(\boldsymbol{\theta} \,|\, Z) \, \mathrm{d}\boldsymbol{\theta} \tag{11}$$

Equation (11) is intractable meaning that approximation methods are required in order to obtain samples from $p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z)$. The proposed approach approximates the two integrals in Eq. (11) using importance sampling; although this can also be thought of as Bayesian model averaging [14, 15] — where a set of models are weighted by their evidence.

Importance sampling approximations for the two integrals in Eq. (11) are performed, setting unnormalised proposal distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_{\delta,j}$. The unnormalised parameter proposal $\tilde{q}_\theta(\boldsymbol{\theta}^{(i)})$ is defined as the approximate posterior distribution from BHM, such that $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \,|\, Z, \boldsymbol{x}_z)$ are N_s re-sampled parameters, all of which are equally likely, i.e. $\tilde{q}_\theta(\boldsymbol{\theta}^{(i)}) \propto 1$. The model discrepancy hyperparameter proposal distribution is chosen to be equal to the prior, i.e. $\tilde{q}_\phi(\boldsymbol{\phi}_{\delta,j}^{(i,k)}) = p(\boldsymbol{\phi}_{\delta,j}^{(i,k)})$; where N_ϕ hyperparameter samples $\boldsymbol{\phi}_{\delta,j}^{(i,k)}$ are obtained for the ith parameter sample $\boldsymbol{\theta}^{(i)}$. These choices of proposal distributions mean that the unnormalised weights become $\tilde{w}_j^{(i,k)} = p(\boldsymbol{z}_j \,|\, \boldsymbol{y}_j^{(i)}, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \boldsymbol{\phi}_{\delta,j}^{(i,k)})$.

For the jth output the method begins by propagating the ith parameter sample $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \,|\, Z, \boldsymbol{x}_z)$ through the GP emulator in order to obtain the samples $\boldsymbol{y}_j^{(i)}$. Subsequently, for the ith parameter sample, hyperparameter samples $\boldsymbol{\phi}_{\delta,j}^{(i,k)} \sim \tilde{q}_\phi(\boldsymbol{\phi}_{\delta,j}^{(i,k)})$ are used in construct N_ϕ GP regression models, mapping from $\boldsymbol{y}_j^{(i)}$ to $\boldsymbol{z}_j$. The marginal likelihood of the ith, kth GP model is equal to $\tilde{w}_j^{(i,k)}$. This process is repeated such that a set of unnormalised weights and unnormalised predictive means $\hat{\boldsymbol{z}}_{*,j}^{(i,k)}$ and covariances $\Sigma_{z*,j}^{(i,k)}$ are obtained. The unconditional



(a) (b)

Figure 2. Representative five storey building structure. Panel (a) show the test setup and panel (b) presents an example of the pseudo-damage, added masses, applied to the first floor.

bias corrected predictions, $p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z)$, are then approximated as Eq. (12) — where the mean and variance are obtained by the laws of total expectation and variance Eqs. (13) and (14).

$$p(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z) \approx \mathcal{N}\left(\mathbb{E}\left(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z\right), \mathbb{V}\left(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z\right)\right) \tag{12}$$

$$\mathbb{E}\left(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z\right) = \sum_{i=1}^{N_s} \sum_{k=1}^{N_\phi} w_j^{(i,k)} \hat{\boldsymbol{z}}_{*,j}^{(i,k)} \tag{13}$$

$$\mathbb{V}\left(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z\right) = \sum_{i=1}^{N_s} \sum_{k=1}^{N_\phi} w_j^{(i,k)} (\Sigma_{z*,j}^{(i,k)} + \hat{\boldsymbol{z}}_{*,j}^{(i,k)} \hat{\boldsymbol{z}}_{*,j}^{(i,k)\mathsf{T}}) - \mathbb{E}\left(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z\right) \mathbb{E}\left(\boldsymbol{z}_{*,j} \,|\, \boldsymbol{x}_*, Z, \boldsymbol{y}_j, \boldsymbol{x}_z\right)^{\mathsf{T}} \tag{14}$$

This process is summarised in Algorithm 2. It is noted that other importance sampling based approaches to marginalising the hyperparameters from a GP are adaptive importance sampling [16], where the proposal is iterative amended in order to improve convergence, and Sequential Monte Carlo (SMC) [17]. These techniques could be implemented to provide faster convergence of the approximations.

4. Representative Five Storey Building Structure

Calibration of five bending modes of a representative five storey building structure was performed using BHM in conjunction with the proposed model discrepancy importance sampling approach. Modal testing was performed on the structure made from aluminium 6082 subject to different pseudo-damage extents as shown in Fig. 2. These pseudo-damage extents were added masses $m = \{0, 0.1, \ldots, 0.5\}$kg fixed to the first floor of the structure demonstrated in Fig. 2b. The structure was excited with 409.6Hz bandwidth Gaussian noise via an electrodynamic shaker, with sample rate and sample time chosen to allow a frequency resolution of 0.05Hz. Accelerometers were placed at each of the five floors in order to obtain the first five bending modes. 40 averages were acquired for each measurement and ten repeats were performed for each damage extent, in order to obtain an understanding of the underlying modal frequency distributions.

The observational data $z(\boldsymbol{x}_z)$ used within the calibration process were the mean natural frequencies when $\boldsymbol{x}_z = \{0, 0.3, 0.5\}$kg. The unseen validation set were the full repeat measurements of $z(\boldsymbol{x}_z)$ as well as those from the $\{0.1, 0.2, 0.4\}$kg pseudo-damage extents, with the inputs collectively denoted as $\boldsymbol{x}_*$.

Parameter		**Lower Bound**	**Upper Bound**
Elastic Modulus	E	63.9GPa	78.1GPa
Poisson's Ratio	ν	0.297	0.363
Density	ρ	2493kg/m^3	3047kg/m^3

Table 1. The prior parameter bounds for BHM on the five storey representative building structure.

Uncertainty		ω_1	ω_2	ω_3	ω_4	ω_5
Observational	V_o	3×10^{-5}	0.02	0.09	0.05	0.01
Model Discrepancy	V_m	1.5	0.01	0.01	1	1

Table 2. The process uncertainties defined in the implausibility measure utilised for performing BHM on the five storey representative building structure.

The simulator $\eta(\boldsymbol{x}, \boldsymbol{\theta})$ was a modal Finite Element Analysis (FEA) model where the five bending natural frequencies were extracted as a set of outputs $\boldsymbol{y}$. Evaluations of the simulator were acquired for the six damage extents $\boldsymbol{x} = \{0, 0.1, \ldots, 0.5\}$kg and a range of parameter $\boldsymbol{\theta}$ values within a set of prior bounds; set as $\pm 10\%$ of typical material properties for aluminium 6082 as shown in Table 1. Simulator runs for parameter combinations determined by a fifty point, three dimensional GMLHD, were implemented as training data for five independent GP emulators, with a separate ten point three dimensional GMLHD used to generate validation data.

4.1. Bayesian History Matching

BHM was implemented using GMLHDs to provide training data for five independent GP emulators. Each emulator was constructed from linear mean and Matérn 3/2 covariance functions. Exploration of the parameter domain was performed via $100,000$ samples from uniform distributions over the bounds. A multivariate implausibility (Eq. (4)) was implemented, with the non-implausibility criteria being when the maximum multivariate implausibility for all five outputs (the five natural frequencies) was less than the 99% quantile for a three degree of freedom χ^2-distribution (reflecting the size of $\boldsymbol{x}_z$). The observational $V_{o,j}$ and model discrepancy $V_{m,j}$ uncertainties, set for the first BHM wave are displayed in Table 2 and were estimated from the experimental output variance for the training inputs and from the modeller's judgement respectively.

The stopping criteria was met after one wave as the code uncertainty for each of the five emulators had an order of magnitude $\approx 10^{-4}$. This low level of code uncertainty indicates that the emulators had captured the simulator behaviour well with diagnostic checks [18] evidencing that the emulators were valid. After the first wave a non-implausible space $\approx 2.3\%$ of the original space was identified.

When the stopping criteria had been met approximate posterior densities can be formed using importance sampling and re-sampling. A Gaussian proposal distribution with $\kappa = 2$ was used to generate $100,000$ samples with which to assess the normalised weights. $100,000$ samples were subsequently obtained by re-sampling the distribution. Figure 3a presents the marginal and joint pairwise posterior distributions, with a linear relationship between low density and high elastic modulus values and a relatively insensitive effect from Poisson's ratio in the pairwise joint distributions. The marginal posterior distribution for each parameter show bi-modal distributions. The pairwise joint posteriors indicate that these modes correspond to opposite ends of the marginal distribution, for example the elastic modulus modal value around 72GPa corresponds to density mode around 2500kg/m^3 and the two Poisson's ratio modes around 0.3

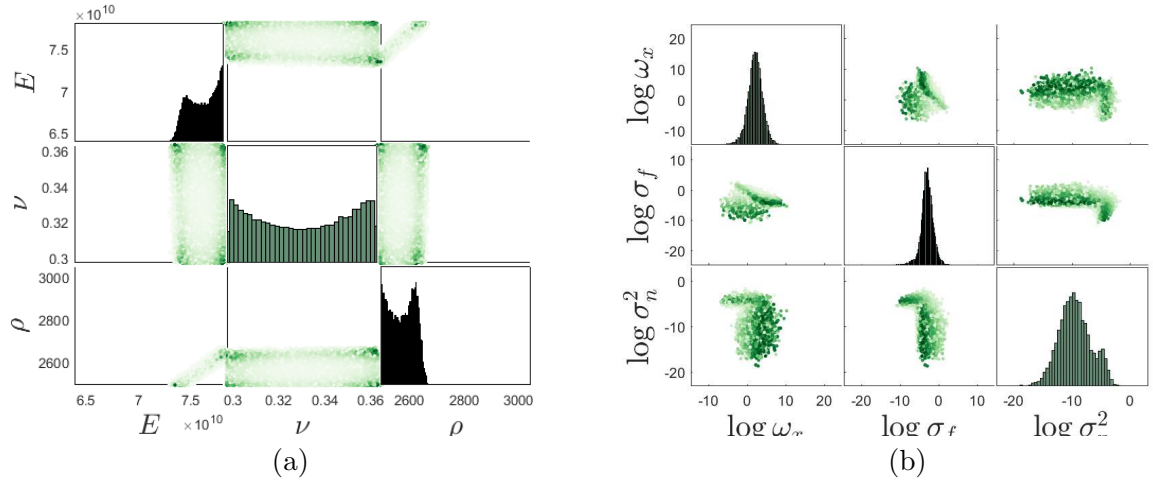


Figure 3. Marginal and pairwise joint posterior distributions for: (a) the parameters identified by the first wave of BHM and (b) the hyperparameters for the fifth natural frequency; where a darker shade represents a higher probability.

and 0.36. These results show the methods ability to account multi-modal behaviour within the defined parameter domain.

The output distributions for each of the five natural frequencies were obtained via Monte Carlo sampling the posterior parameter distribution. $1,000$ samples were taken from the re-sampled parameter posterior distributions and propagated through each of the five emulators in order to obtain realisations of the output distributions. As the code uncertainty across all emulators was extremely low $\approx 10^{-4}$, the emulator mean was taken as deterministic. It is noted that if the emulator variances were not several orders of magnitude lower than the combined observational and model discrepancy uncertainties to be deemed negligible, posterior sampling of the GP should be implemented (as shown in [19]). The mean predictions of the GP emulators for the 1000 Monte Carlo realisations are presented in Fig. 4 against the observational data used within BHM $z(\boldsymbol{x_z})$ with $\pm c_\sigma(V_{o,j} + V_{m,j})$ bounds, where c_σ is the standard deviation associated with 99% probability mass of a standard normal (assuming output distributions to be approximately Gaussian). Figure 4 demonstrates that all five outputs are within the defined uncertainty bounds. However, large discrepancies between the experimental observations and simulator (represented by the five emulator's mean predictions) occur, especially for the first and fifth natural frequencies. This illustrates that the simulator has model form errors, that would lead to incorrect parameter inference if model discrepancy was not considered in the calibration process.

4.2. Model Discrepancy Inference

The proposed importance sampling methodology was applied to the BHM outputs. Here $N_\phi = 100$ and $N_s = 1000$, with the model discrepancy GPs being zero mean and Matérn $3/2$ covariance functions. The hyperparameter priors were Gaussian distributed — $\log \omega_{x,j} \sim \mathcal{N}(0, 6)$ (covariance roughness parameter), $\log \sigma^2_{fj} \sim \mathcal{N}(\mathbb{V}(\boldsymbol{z}_j), 4)$ (covariance signal variance) and $\log \sigma^2_{nj} \sim \mathcal{N}(V_{o,j} - 5, 6)$ (the noise variance) — stating that a low noise and smooth model discrepancy solution is expected.

Figure 5 presents the predictive output distributions after the importance sampling approach. This demonstrates that the model discrepancy has been inferred correctly for the second to fifth natural frequencies. On the other hand the first natural frequency fails to fit the 0.1 and 0.2kg

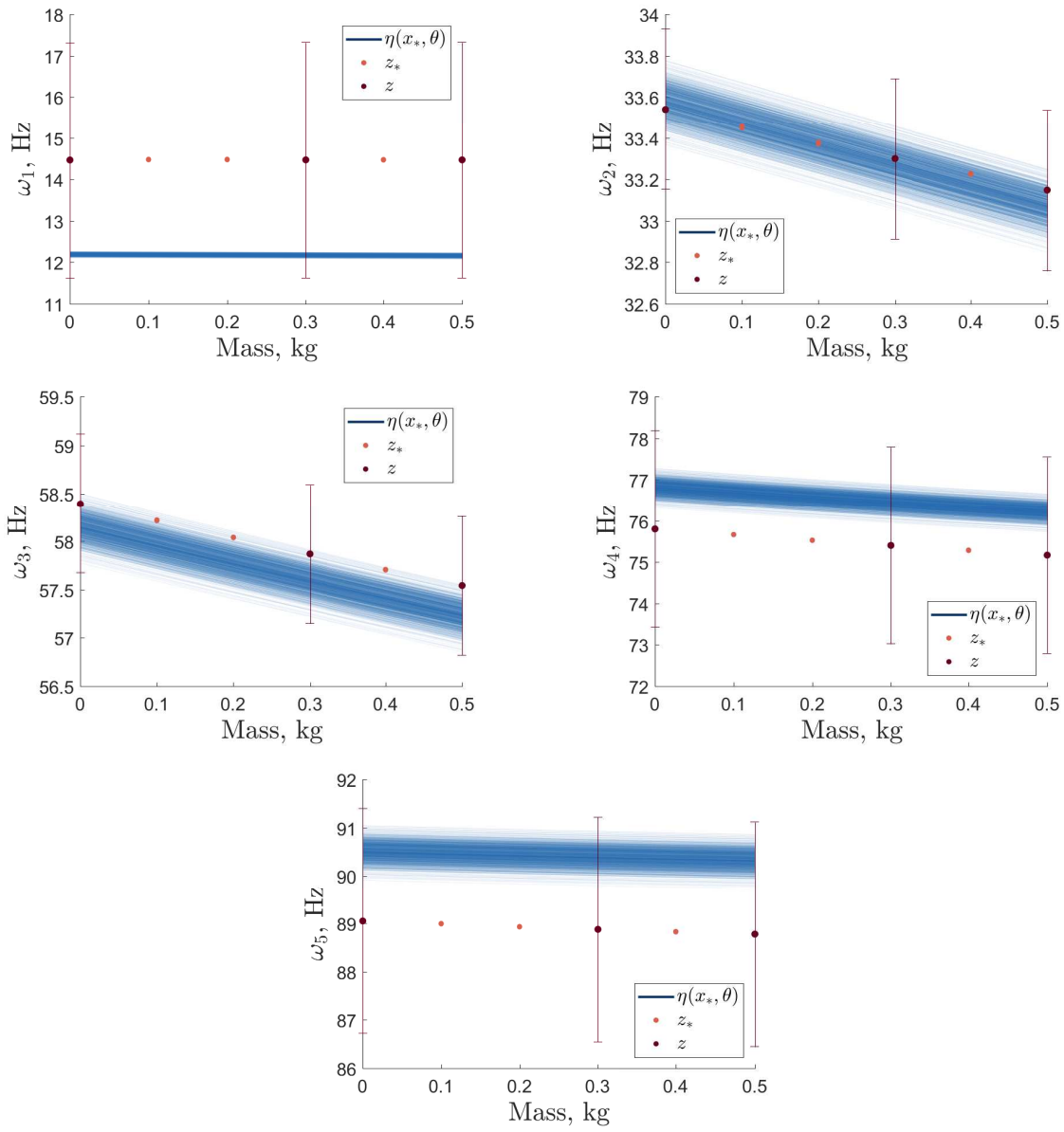


Figure 4. 1000 samples of the BHM predictive outputs, $p(\boldsymbol{y}^{(i)}_{*,j} \,|\, \boldsymbol{x}_*, \boldsymbol{y}_j, \boldsymbol{x}_z, \boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\phi}}_{\eta,j})$ given $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \,|\, Z, \boldsymbol{x}_z)$. The error bars indicate the defined model discrepancy and observation variances.

mass states. This is likely due to a lack of information about these point in the training set. It would be expected that with inclusion of these data points predictions would be improved. The posterior distribution over the hyperparameters is estimated from the weights; Fig. 3b demonstrate these distributions for the fifth natural frequency, where the type I and II maximum likelihoods are clearly indicated within the pairwise joint densities, where each refers to the noise and roughness invariant solutions.

Validation metrics were applied to the output predictions, with the Normalised Mean Squared Error (NMSE)s between the mean predictions and observational data being 145.11, 0.51, 0.25, 0.34 and 0.94. These indicate very good mean fits, apart from the first natural frequency.

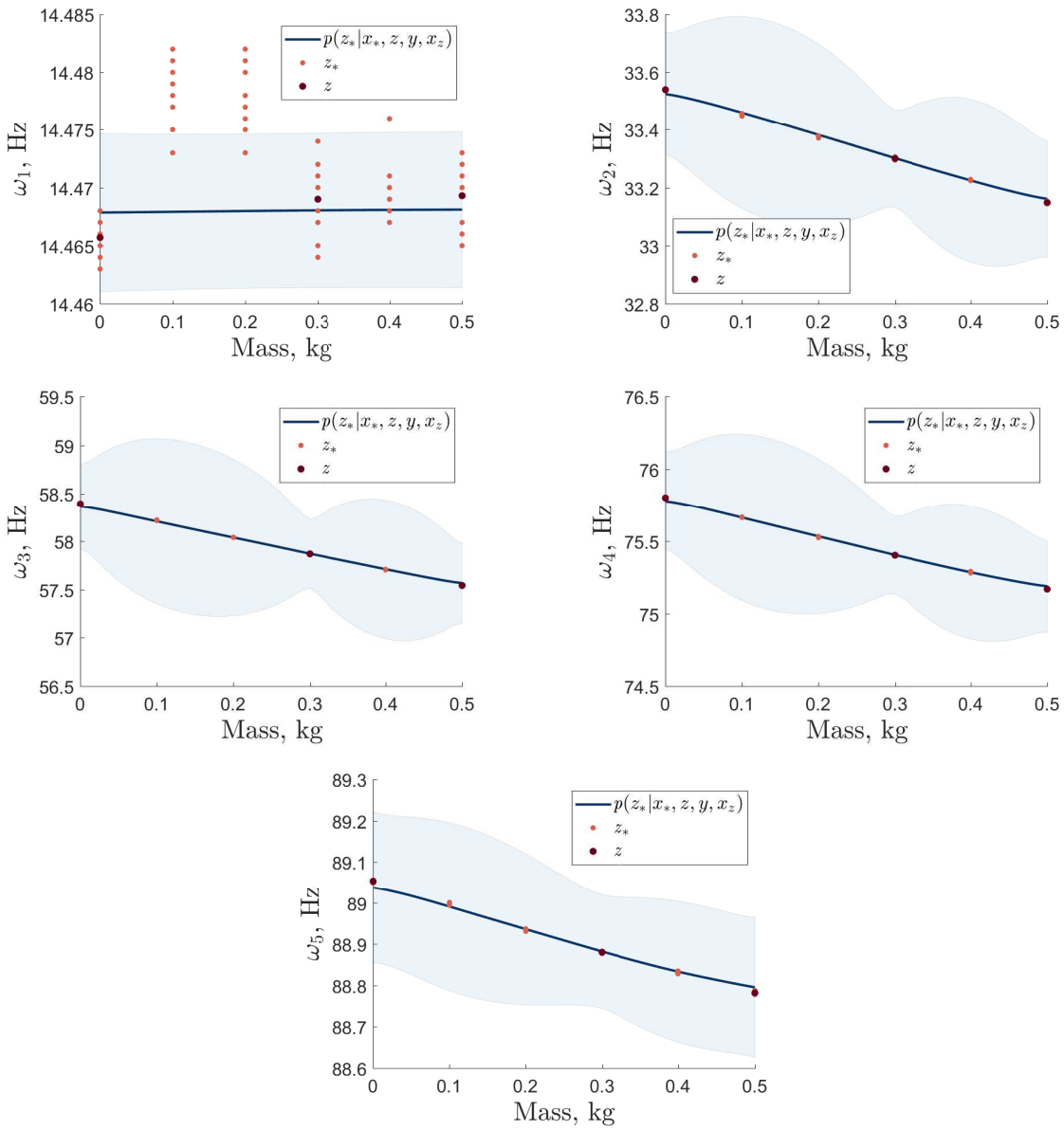


Figure 5. BHM predictive outputs with the marginalisation of model discrepancy and GP hyperparameters via importance sampling — prior proposal.

Statistical distance metrics that assess the distance between two distributions were quantified. Firstly, the area metric, the area between the two Cumulative Density Function (CDF)s [20], was applied as shown in Fig. 6a. This metric showed small frequency distances across all predictions, with the 0.1 and 0.2kg mass states for the first natural frequency being the largest. Secondly, Hellinger distances, the integral of the L_2-norm between two Probability Density Function (PDF)s [21], were applied, Fig. 6b. The units of this metric are normalised distances and show relatively large distances for all but the first natural frequency's 0, 0.3, 0.4 and 0.5kg predictions. This is due to a larger predictive variances than observational variances, likely caused by the high level of uncertainty due to both the parameter and model discrepancy inferences.

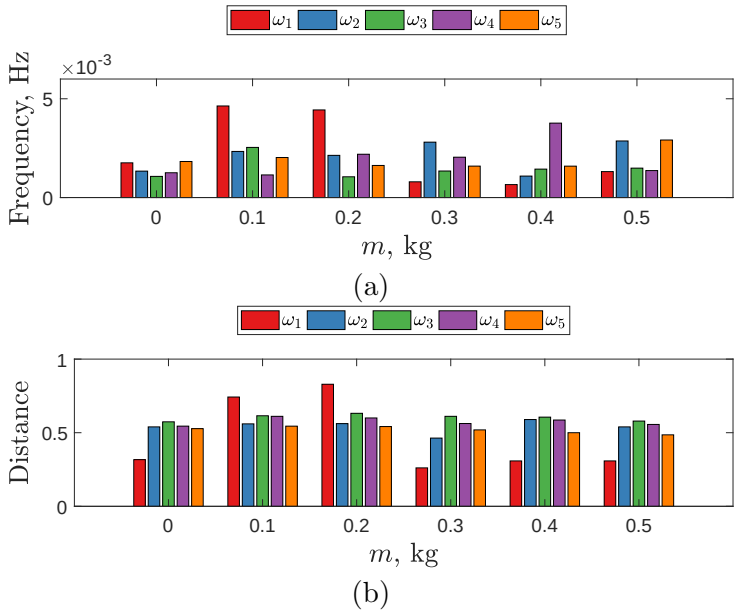


Figure 6. Statistical distances applied to the bias corrected predictions. Panel (a) are the area metrics when compared to empirical ten point observational CDFs. Panel (b) are the Hellinger distances when compared to KDEs of the observational data. Both metrics are calculated via numerical integration.

5. Conclusions

Model discrepancy must be accounted for, and inferred, when calibrating structural dynamics simulators. This paper presented a BHM methodology for calibrating simulators in the presence of additive, uniform model discrepancy. In addition, an importance sampling based approach has been developed and demonstrated as an effective method for inferring the uncertainty associated with the model discrepancy functional form.

This technique has been applied to a representative five storey building structure where it has been shown to be effective for both calibration and inference of model discrepancy. The importance sampling methodology captured the model discrepancy improving predictive quality for the majority of outputs. However, for the first natural frequency, where the training data was not representative of the functional behaviour, the predictive distributions failed to capture two input points. This would be improved by more informative training data. In addition, due to the high level of uncertainty, the output predictions had a larger variance than the observations leading to large Hellinger distances. These would be improve by increased training observations, or by improving the model form errors within the simulator.

Lastly, further work should be conducted in improving the computationally efficient in the model discrepancy importance sampling. This would allow fewer GP models to be constructed making the approach more practical for scenarios where more observational data is available. Furthermore multivariate GPs should be investigated as priors for the emulators and model discrepancies. This may improve parameter and model discrepancy inferences when there is dependency between outputs.

References

[1] Peter S Craig, Michael Goldstein, Allan H Seheult, and James A Smith. Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments. In *Lecture Notes in Statistics*, pages 37–93. 1997.

[2] Michael Goldstein and Rui Paulo. External Bayesian Analysis for Computer Simulators. *Bayesian Statistics 9*, (1996), 2012.
[3] Ian Vernon, Michael Goldstein, and Richard Bower. Galaxy Formation: Bayesian History Matching for the Observable Universe. *Statistical Science*, 29(1):81–90, feb 2014.
[4] Ioannis Andrianakis, Ian R. Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, and Richard G. White. Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda. *PLoS Computational Biology*, 11(1), 2015.
[5] I. Andrianakis, I. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White. History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(4):717–740, aug 2017.
[6] Neil R. Edwards, David Cameron, and Jonathan Rougier. Precalibrating an intermediate complexity climate model. *Climate Dynamics*, 37(7-8):1469–1482, 2011.
[7] Daniel Williamson, Michael Goldstein, Lesley Allison, Adam Blaker, Peter Challenor, Laura Jackson, and Kuniko Yamazaki. History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, 41(7-8):1703–1729, 2013.
[8] A O'Hagan and JFC Kingman. Curve Fitting and Optimal Design for Prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42, 1978.
[9] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.*, volume 14. 2004.
[10] Thomas E. Fricker, Jeremy E. Oakley, and Nathan M. Urban. Multivariate Gaussian Process Emulators With Nonseparable Covariance Structures. *Technometrics*, 55(1):47–56, 2013.
[11] K. Worden, G. Manson, and N. R.J. Fieller. Damage detection using outlier analysis. *Journal of Sound and Vibration*, 229(3):647–667, 2000.
[12] Pukelsheim. The three sigma rule. *American Statistician*, 48(2):88–91, 1994.
[13] Holger Dette and Andrey Pepelyshev. Generalized Latin Hypercube Design for Computer Experiments. *Technometrics*, 52(4):421–429, nov 2010.
[14] Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.
[15] Le Bao, Tilmann Gneiting, Eric P. Grimit, Peter Guttorp, and Adrian E. Raftery. Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction. *Monthly Weather Review*, 138(5):1811–1821, 2010.
[16] Dejan Petelin, Gasperin Matej, and Vaclav Smidl. Adaptive Importance Sampling for Bayesian Inference in Gaussian Process models. pages 5011–5016, 2014.
[17] Andreas Svensson, Johan Dahlin, and Thomas B. Schön. Marginalizing Gaussian Process Hyperparameters using Sequential Monte Carlo. pages 4–7, 2015.
[18] Leonardo S. Bastos and Anthony O'Hagan. Diagnostics for Gaussian Process Emulators. *Technometrics*, 51(4):425–438, 2009.
[19] Philipp Hennig and Christian J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13:1809–1837, dec 2012.
[20] Scott Ferson, William L. Oberkampf, and Lev Ginzburg. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32):2408–2430, 2008.
[21] P Gardner, C Lord, and R J Barthorpe. An Evaluation of Validation Metrics for Probablistic Model Outputs. In *Proceedings of the ASME 2018 Verification and Validation Symposium*, 2018.