Mechanical Systems and Signal Processing 143 (2020) 106828

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Bayesian history matching for structural dynamics applications

P. Gardner *, C. Lord, R.J. Barthorpe

Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, United Kingdom

ARTICLE INFO

Article history:
Received 19 September 2019
Received in revised form 23 December 2019
Accepted 19 March 2020

Keywords:
Bayesian history matching
Model discrepancy
Calibration
Parameter estimation
Uncertainty quantification

ABSTRACT

Computer models provide useful tools in understanding and predicting quantities of interest for structural dynamics. Although computer models (simulators) are useful for a specific context, each will contain some level of model-form error. These model-form errors arise for several reasons e.g., numerical approximations to a solution, simplifications of known physics, an inability to model all relevant physics etc. These errors form part of *model discrepancy*; the difference between observational data and simulator outputs, given the 'true' parameters are known. If model discrepancy is not considered during calibration, any inferred parameters will be biased and predictive performance may be poor. Bayesian history matching (BHM) is a technique for calibrating simulators under the assumption that additive model discrepancy exists. This 'likelihood-free' approach iteratively assesses the input space using emulators of the simulator and identifies parameters that could have 'plausibly' produced target outputs given prior uncertainties. This paper presents, for the first time, the application of BHM in a structural dynamics context. Furthermore, a novel method is provided that utilises Gaussian Process (GP) regression in order to infer the missing model discrepancy functionally from the outputs of BHM. Finally, a demonstration of the effectiveness of the approach is provided for an experimental representative five storey building structure.

1. Introduction

Calibration of computer models (herein defined as simulators) is often an important aspect of creating predictions that adequately match observational data. However, simulators often contain model-form errors from various sources, such as an absence and/or simplification of the physics and approximations in solution techniques. These errors form part of the term *model discrepancy*, which is defined as the mismatch between observational data and the simulator output when the 'true' parameters are known. If a mechanism that accounts for model discrepancy is not included in the parameter estimation approach, any parameters identified will be biased and may provide poor predictions, especially when extrapolating [1,2]. Consequently, Bayesian History Matching (BHM) has been developed as a methodology for calibrating simulators under the assumption that model discrepancy exists and can be treated as uncertain.

History matching as a term originates from the oil industry and describes methods that find parameters of simulators where the outputs closely match data from historical reservoir production. Many of these techniques within the literature, such as those reviewed by Oliver and Chen [3], are similar to classical model updating techniques well-established within structural dynamics [4,5]. Nonetheless, Craig et al. adapted the idea of history matching from the oil industry and outlined

* Corresponding author.
E-mail address: p.gardner@sheffield.ac.uk (P. Gardner).

https://doi.org/10.1016/j.ymssp.2020.106828
0888-3270/

an approximate Bayesian methodology that searched for all, rather than a single parameter match [6]. This category of approaches was defined as Bayesian History Matching.

BHM has been implemented across a wide variety of applications, from its origins in oil reservoir modelling [6], to understanding Galaxy formation [7–9], complex social models of HIV transfer in populations [10,11] and climate science [12,13]. The method seeks to discard parts of the parameter space that were unlikely to produce outputs that match the observational data, given that observational and model discrepancy uncertainties exist. A key benefit of the approach is in gaining an understanding about parameters in complex computer models where model discrepancy is present. In addition, Andrianakis et al. in [10] discuss utilising the method as a pre-calibration step before applying fully Bayesian analysis, such that the parameter domain is well understood, and informative priors identified. However, none of the applications of BHM within the literature apply the technique to a structural dynamic context, nor do they seek to identify the functional form of the model discrepancy after the parameter distributions are estimated. Correspondingly, one of the primary novel contributions of this paper is in providing a methodology for inferring the functional form of the model discrepancy after calibration via BHM. This is performed by using the maximum a *posteriori* (MAP) estimate of the inferred parameter posterior distribution such that Gaussian Process (GP) regression models can be inferred to map between the simulator output and a set of training observational data. This novel combined approach provides an alternative method to conventional Bayesian calibration methods that consider model discrepancy within the parameter inference process [1,2,14–16].

Another benefit of BHM is that the approach is 'likelihood-free' meaning that input and output combinations can be removed and added iteratively without invalidating the analysis. 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; which are often present in current Bayesian approaches to calibration that consider model discrepancy [1,2,14–16]. Moreover, by decoupling the parameter and model discrepancy estimation problem, BHM offers a further mechanism to prevent against non-identifiability issues and non-physical inferences. The 'likelihood-free' approach also provides benefits in allowing approximate Bayesian solutions to be identified when a valid likelihood is not possible to obtain.

The technique is designed to be computationally competitive, when compared to Bayesian sampling methods such as Markov Chain Monte Carlo (MCMC), and therefore practical for complex simulators. By construction, BHM utilises emulators—computationally efficient surrogate models—such that the parameter domain can be explored efficiently. Within the literature both Bayes linear techniques [8,9] and Gaussian Process (GP) [10] emulators have been implemented. These choices are popular as both provide estimates of *code uncertainty* [1]—the uncertainty introduced by replacing the simulator with an emulator. Quantifying code uncertainty ensures that regions of the parameter space are not discarded until emulator predictions are sufficiently 'certain'.

The outline of this paper is as follows. Section 2 introduces BHM defining assumptions and aspects of implementation. Following the methodology, a numerical case study is presented in Section 3 such that the effectiveness of BHM can be demonstrated on an example where the true parameters are known. Next, BHM is demonstrated on experimental case study of a representative five storey building structure, where masses are used to simulate pseudo-damage scenarios. This section also outlines and implements the approach for inferring the functional form of the model discrepancy. Finally, conclusions are discussed, highlighting areas for further research.

2. Methodology

BHM is an approximate Bayesian approach for determining whether parameter combinations Θ are 'implausible'; that is to say not likely to have produced known observations $\boldsymbol{z}$. These implausible parameter combinations $\boldsymbol{\theta}_I \in \Theta$ are discarded based on a criteria such that the remaining non-implausible space $\boldsymbol{\theta}_{nI} \in \Theta$ are identified. In terms of a statistical model, BHM aims to calibrate,

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

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 and e respectively, where the simulator, model discrepancy and observational uncertainty are independent. Eq. (1) is similar to that proposed by Kennedy and O'Hagan in [1]; although here model discrepancy is defined as constant and additive with respect to the inputs.

In order to calibrate Eq. (1), BHM explores the parameter space of the simulator iteratively, where each iteration is called a *wave*. During a wave simulator outputs are assessed for various parameter combinations using an *implausibility metric* and discarded if above a threshold T. As the method is required to assess a large parameter space, a computationally efficient emulator is utilised. The technology used in constructing an emulator within BHM must also quantify *code uncertainty* [1]—the uncertainty introduced by replacing the simulator with an emulator. This is important as code uncertainty prevents regions of the parameter space from begin discarded due to poor emulation; instead these are retained until the emulator sufficiently represents the simulator output. For this reason GP emulators are implemented—Bayesian, non-parametric regression models [17,18]—as they will fit known simulator runs exactly (under a no noise assumption) and provide estimations of code uncertainty when predicting away from known simulator runs. It is noted that Bayes linear techniques could also be used, as these also quantify code uncertainty [8,9]. However, these approaches are approximation methods that

update beliefs using linear fitting, and are generally more applicable to scenarios where the simulator space is not well modelled by a GP.

2.1. Gaussian process emulators

A GP emulator is constructed as,

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

where the jth simulator output is modelled as a GP, $\mathcal{GP}_j(\cdot,\cdot)$; taking both $\boldsymbol{x}$ and $\boldsymbol{\theta}$ as its arguments—it is noted that the emulator is a map of both the inputs $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ to the jth simulator output $\eta_j(\boldsymbol{x},\boldsymbol{\theta})$. The GP emulator is fully defined by its mean $m(\cdot)$ and covariance $k(\cdot,\cdot)$ functions, which define the prior belief over the functions that could represent the simulator output. Here the mean function is defined to be linear in the parameters, i.e. $m(\cdot) = H\boldsymbol{\beta}$, where H is comprised of p basis functions $H = (h_1(\cdot),\ldots,h_p(\cdot))$ and there are p coefficients $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$. Is it noted that the basis functions in H should reflect the prior belief about $\eta(\cdot,\cdot)$, where typically no more than a constant or linear mean can be assumed *a priori*. The covariance function states the prior assumption about how smooth the simulator output is. It defines the correlation between any two points in a Reproducing Kernel Hilbert Space (RKHS) and is dependent on some set of hyperparameters $\boldsymbol{\phi}_\eta$, i.e. $K = k((\cdot,\cdot);\boldsymbol{\phi}_\eta)$. One example of a covariance function is the squared exponential covariance function,

$$k((\boldsymbol{x},\boldsymbol{\theta}),(\boldsymbol{x}',\boldsymbol{\theta}')) = \sigma_\eta^2 \exp\left(-(\boldsymbol{x}-\boldsymbol{x}')^T\Omega_x(\boldsymbol{x}-\boldsymbol{x}')\right)\exp\left(-(\boldsymbol{\theta}-\boldsymbol{\theta}')^T\Omega_\theta(\boldsymbol{\theta}-\boldsymbol{\theta}')\right) \tag{3}$$

where the hyperparameters $\boldsymbol{\phi}_\eta$ for the squared exponential are two diagonal matrices of roughness parameters i.e. $\Omega = \mathrm{diag}(\omega_1,\ldots,\omega_D)$ and a signal variance σ_η^2 (the reader is referred to [18] for more examples and definitions of covariance functions). By forming a joint prior over the training (denoted y) and testing (denoted $*$) points, standard Gaussian conditionals can be used to obtain the predictive distribution,

$$p(\boldsymbol{\eta}_*|\boldsymbol{x}_*,\boldsymbol{\theta}_*,\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{N}(\mathbb{E}(\boldsymbol{\eta}_*),\mathbb{V}(\boldsymbol{\eta}_*)) \tag{4a}$$

$$\mathbb{E}(\boldsymbol{\eta}_*|\boldsymbol{x}_*,\boldsymbol{\theta}_*,\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = H_*\boldsymbol{\beta} + K_{*,y}K_{y,y}^{-1}(\boldsymbol{y}-H_y\boldsymbol{\beta}) \tag{4b}$$

$$\mathbb{V}(\boldsymbol{\eta}_*|\boldsymbol{x}_*,\boldsymbol{\theta}_*,\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = K_{*,*} - K_{*,y}K_{y,y}^{-1}K_{y,*} \tag{4c}$$

where $\boldsymbol{\eta}_*$ are predictions of the simulator function at the test points and Eq. (4c) quantifies the code uncertainty. Due to the focus of this paper being BHM and not GP regression, the reader is referred to [17–20] for more mathematical details.

It is noted that the computation cost of training a GP emulator is $\mathcal{O}(n^3)$ (where n is the number of training points in $K_{y,y}$). It is expected that in applications involving computationally expensive simulators the computational cost in training a GP emulator will be insignificant compared to obtaining simulator evaluations. Nonetheless the computational cost in training a GP emulator can become problematic when the mapping from inputs $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ to the jth simulator output $\eta_j(\boldsymbol{x},\boldsymbol{\theta})$ is complex compared to the prior (fully specified by $m(\cdot)$ and covariance $k(\cdot,\cdot)$), or when the parameter and input spaces are significantly high dimensional, and therefore a large number of simulator evaluations are required to adequately train the GP. However, in these scenarios sparse GP approximations can be utilised, where the computational cost of training the GP emulator reduces from $\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2)$, where typically $m \ll n$ and is a number of inducing points (for more details on sparse GP emulators see [21]).

2.2. Implausibility metric

The implausibility metric $I(\cdot,\cdot)$ [10,11], used to determine whether a given parameter combination was likely to have produced the observed output, incorporates the emulators within its formulation,

$$I_j(\boldsymbol{x},\boldsymbol{\theta}) = \frac{|z_j(\boldsymbol{x}) - \mathbb{E}_j(\boldsymbol{\eta}_*|\boldsymbol{x}_*,\boldsymbol{\theta}_*,\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi})|}{\left(V_{o,j}+V_{m,j}+V_{c,j}(\boldsymbol{x},\boldsymbol{\theta})\right)^{1/2}} \tag{5}$$

where V_o, V_m and $V_c(\boldsymbol{x},\boldsymbol{\theta})$ are the variances associated with the observational, model discrepancy and code uncertainties; $V_c(\boldsymbol{x},\boldsymbol{\theta}) = \mathbb{V}_j(\boldsymbol{\eta}_*|\boldsymbol{x}_*,\boldsymbol{\theta}_*,\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi})$. Eq. (5) is essentially the distance between observational data and emulator mean predictions weighted by the uncertainties in the processes. Trivially, in the case where simulator runs are computationally cheap, the emulator mean can be replaced with the simulator output and the code uncertainty term removed.

The implausibility metric requires specification of the observational and model discrepancy uncertainties, V_o and V_m. The observational uncertainty V_o can often be estimated from expert knowledge and from the observational data. Model discrepancy uncertainty V_m can be more challenging to define, but may be elicited from expert judgement. The likelihood-free property of BHM means that the model discrepancy uncertainty can be refined during each wave. This means that sensitivity analysis of the effect of V_m can be performed during a wave to understand changes in rejection rates and help improve

its specification. Observational and model discrepancy uncertainties can be dependant on both inputs $\boldsymbol{x}$ and outputs $z_j(\boldsymbol{x})$, i.e. $V_{o,j}(\boldsymbol{x})$ and $V_{m,j}(\boldsymbol{x})$, if input dependent hetroscedastic noise or model discrepancy are hypothesised.

The implausibility metric presented in Eq. (5) provides a quantity for every parameter combination $\boldsymbol{\theta}$, input $\boldsymbol{x}$ and output $\boldsymbol{y}$. 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 [10]. Firstly, a maximum implausibility can be formed,

$$I_{max}(\boldsymbol{\theta}) = \underset{j}{\operatorname{argmax}} \left(\underset{\boldsymbol{x_i}}{\operatorname{argmax}} I_j(\boldsymbol{x}, \boldsymbol{\theta}) \right) \tag{6}$$

whereby the worst case for a given parameter combination is used. Another approach is to form a multivariate implausibility metric for either the inputs,

$$I_{multi}(\boldsymbol{\theta})_j = \left(z_j(\boldsymbol{x}) - \mathbb{E}(\mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta}))\right)^T \left(V_{o,j} + V_{m,j} + V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta})\right)^{-1} \left(z_j(\boldsymbol{x}) - \mathbb{E}(\mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta}))\right) \tag{7}$$

or outputs,

$$I_{multi}(\boldsymbol{x}, \boldsymbol{\theta}) = \left(z_j(\boldsymbol{x}) - \mathbb{E}(\mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta}))\right)^T \left(V_{o,j} + V_{m,j} + V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta})\right)^{-1} \left(z_j(\boldsymbol{x}) - \mathbb{E}(\mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta}))\right) \tag{8}$$

which is equivalent to taking the Mahalanobis distance; assessing the Euclidean distance of the principal components (standard practice in outlier analysis [22]). Again a maximum can be taken over either Eq. (7), (8) to collapse the metric to a single value for each parameter combination.

2.3. Decision threshold

In order to decide which parts of the parameter space to exclude, a threshold T is placed on the implausibility metric. Large implausibilities (for each formulation) indicate a parameter set was very unlikely to have produced an output that matched the observational data, given the included uncertainties. Using this knowledge, a rejection criteria can be formed for a particular parameter combination $\boldsymbol{\theta}$,

$$I(\boldsymbol{\theta}) \begin{cases} \leqslant T & \text{if } \boldsymbol{\theta} \in \boldsymbol{\theta}_{nI}, \\ > T & \text{if } \boldsymbol{\theta} \in \boldsymbol{\theta}_{I} \end{cases} \tag{9}$$

where determining the value of T changes based on the type of implausibility metric considered.

Andrianakis et al. [10] state that a sensible threshold T for single $I_j(\boldsymbol{x}, \boldsymbol{\theta})$ or maximum $I_{max}(\boldsymbol{\theta})$ implausibilities (where the maximum is of a single implausibility set) can be determined by Pukelsheim's 3σ rule [10]. The rule states that any continuous unimodal distribution will contain at least 99.5% of probability mass within three standard deviations away from the mean [23]. For multivariate implausibilities the threshold T can be set as a high percentile (e.g. $\alpha > 95\%$) from a chi-squared distribution with either j, or the input size of $\boldsymbol{x}$, degrees of freedom [10], i.e. $T = \mathbf{F}^{-1}_{\chi^2}(\alpha)$ the output from a chi-squared quantile function (inverse Cumulative Density Function (CDF)). This can be thought of as performing a frequentist hypothesis test on the parameter combination, using a chi-squared (χ^2) test.

2.4. Parameter domain exploration

During each wave BHM explores the parameter space using the implausibility metric and threshold. This requires sampling the parameter domain using a design of experiments, and then running the simulator such that outputs can be obtained with which to form the emulator; a Latin hypercube-based approach is a natural choice for constructing the emulator. In this scenario the initial parameter space bounds are used in conjunction with a simulator budget to construct a Latin hypercube design. In this paper Generalised Maximum Latin Hypercube (GMLHC) designs are used as this approach has been shown to reduce the code uncertainty in GP emulators at the bounds of the design [24]; this is particularly useful in BHM, as decisions about whether parameters near the bounds are implausible can be made without high emulator uncertainty. At each wave a new GMLHC can be formed such that the non-implausible space from the last wave can be interrogated.

Once an emulator is constructed from a cost-effective number of simulator runs it is deployed in sampling the parameter domain. In this paper parameter combinations are sampled from a uniform distribution bounded by the initial parameter domain—effectively defining a uniform prior over the space. Emulator predictions are made for each of these samples and a decision made regarding whether they should be discarded.

2.5. Algorithm

The algorithm for performing BHM is stated in Algorithm 1. Two stopping criteria are constructed based on the following outcomes: all the space is deemed implausible; or the code uncertainty in the non-implausible region is less than the remaining uncertainties, i.e. $V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta}_{nI}) < V_{o,j} + V_{m,j}$. This second stopping criteria 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.

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
$\boldsymbol{\theta}_s^k \sim \mathcal{U}(\min(\boldsymbol{\theta}^k), \max(\boldsymbol{\theta}^k))$ ▷n samples of the parameter space
for j = 1: No. of outputs **do**
  Train and validate $\mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta}^k)$ ▷Train and validate emulators
  $\left[\mathbb{E}(\boldsymbol{\eta}_*), V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta}_s^k)\right] = \mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta}_s^k)$ ▷Predict at n samples of $\boldsymbol{\theta}^k$
  Calculate $I_j(\boldsymbol{x}, \boldsymbol{\theta}_s^k)$ ▷Assess implausibility of samples
end for
Calculate $I_{max}(\boldsymbol{\theta}_s^k)$
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

To illustrate BHM, Algorithm 1 is applied to a simple numerical example (where the sampling stage is replaced with a uniform grid). In the example a simulator constructed from Eq. (10a) models the experimental observation z, which is obtained from the 'true' process with noise, stated in Eq. (10b); where $e \sim \mathcal{N}(0, 0.05)$. The observation $z(0.9) = 3.39$ has observational and model discrepancy uncertainties, $V_o = 0.05$ and $V_m = 0.04$ (estimated from the residual variance $\mathbb{V}((z - e) - y)$).

$$y = \eta(\theta) = 5.5(0.15\cos(2\pi \times 0.75\theta) + 1.25\sin(2\pi \times 0.1\theta)) \tag{10a}$$

$$z(\theta) = y(\theta) - 0.3\sin(2\pi \times 0.15\theta) + e \tag{10b}$$

Fig. 1a presents the experiential data point $z(0.9) = 3.39$ with $\pm\sqrt{V_o}$ intervals (shaded region) against the simulator and bias-corrected outputs (i.e. $z - e$) across the parameter space $\theta_s = \{-0.5, 0.005, \ldots, 5\}$ where a budget of four simulator evaluations have been performed in a space-filling manner $\theta^1 = \{0.75, 1.25, 1.75, 2.25\}$. The observation $z = 3.39$ can be formed from two parameter 0.90 and 4.23 indicated by the cross-over in Fig. 1a.

BHM was performed following Algorithm 1 with a simulator evaluation budget of four (for each space-filled design in wave k) where the single implausibility metric $I(\theta)$ and threshold $T = 3$ are implemented. The emulator for each wave was constructed from constant mean and squared exponential covariance functions. The first, second and fourth waves are shown in Fig. 1.

In the first wave (Fig. 1b) the emulator predictions are most uncertain outside of θ^1 leading to these regions being classified as non-implausible. It can also be seen that the initial known simulator runs are deemed implausible, which can be visually confirmed as they are not within the remaining uncertainty bounds $z \pm \sqrt{V_o + V_m}$. Between these known simulator runs the code uncertainty increases leading to the parameters, around 1 and 2, being classed as non-implausible. By the second wave (Fig. 1c) additional simulator runs mean that the code uncertainty in the [0.75 2.25] interval are reduced below the remaining uncertainties and all judged as implausible. Simulator runs at the parameter bounds 'pin' the code uncertainty removing the domain edges as implausible. By the final wave (Fig. 1d where $k = 4$) the code uncertainty has reduced across the space and is lower than the remaining uncertainties in the non-implausible region. The non-implausible set $\boldsymbol{\theta}_{nI}$ at this wave clearly contain two regions around the solution 0.90 and 4.23.

2.6. Approximate posterior sampling

Once the final wave has identified a non-implausible parameter region, importance sampling [10] can be used to obtain an approximation of the posterior distribution $p(\boldsymbol{\theta}|z)$. The approximate posterior is formed from the ratio,

$$p(\boldsymbol{\theta}|z) \approx \frac{w^{un}(\boldsymbol{\theta}_q)}{\frac{1}{n}\sum_{i=1}^{n} w^{un}(\boldsymbol{\theta}_q)} \tag{11}$$

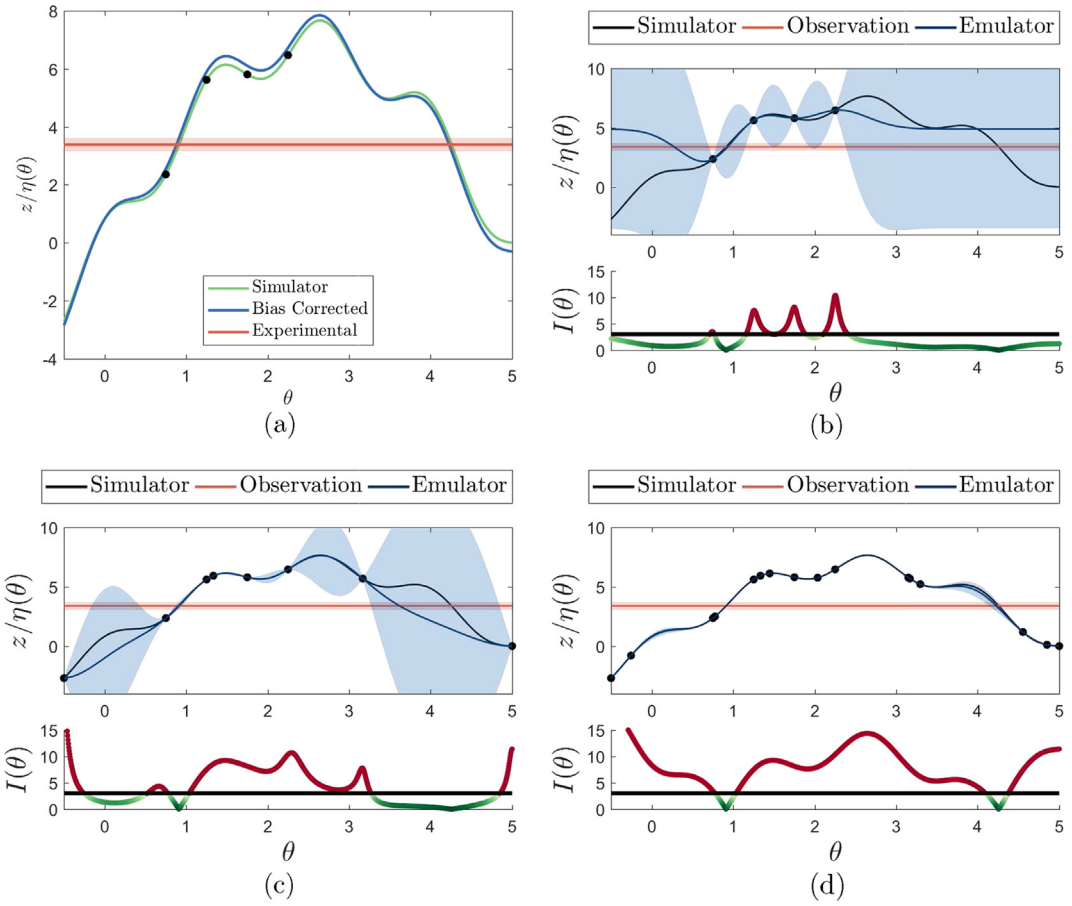

Fig. 1. BHM numerical example. Panel (a) presents the simulator, model discrepancy and observational data (where the shaded region is $\pm\sqrt{V_o + V_m}$) and the initial simulator are (•). Panels (b), (c) and (d) are BHM waves $k = 1, 2, 4$. The top figure in these panels show the observational data with $\pm\sqrt{V_o + V_m}$ shaded region against the simulator and emulator predictions (where the shaded regions indicates $\pm 3\sigma$), trained using the simulator runs $\eta(\boldsymbol{\theta}^k)$ (•). The bottom figures in these panels present the implausibility $I(\boldsymbol{\theta}_s^k)$ against the threshold $T = 3$, where green regions are non-implausible and red implausible.

where $w^{un}(\cdot)$ are a set of un-normalised weights and $\boldsymbol{\theta}_q$ are samples from a proposal distribution $q(\boldsymbol{\theta}_q)$. The un-normalised weights $w^{un} = p(z|\boldsymbol{\theta}_q)p(\boldsymbol{\theta}_q)/q^{un}(\boldsymbol{\theta}_q)$ are the probability of each sample $\boldsymbol{\theta}_q \sim q^{un}$.

However, BHM is 'likelihood-free', and therefore an approximation of the likelihood must be formed. In this paper the likelihood is approximated as,

$$p(z|\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}) = \prod_{j=1}^{M} \mathcal{N}(z(\boldsymbol{x})|\mathbb{E}_j(\mathcal{GP}_j(\boldsymbol{x}, \boldsymbol{\theta})), V_j(\boldsymbol{x}, \boldsymbol{\theta})) \tag{12}$$

which is the product of multivariate Gaussian distributions over $z(\boldsymbol{x})$ for the set of inputs $\boldsymbol{x}$, where $V_j(\boldsymbol{x}, \boldsymbol{\theta}) = V_{o,j} + V_{m,j} + V_{c,j}(\boldsymbol{x}, \boldsymbol{\theta})$. This approximation reflects the form of the implausibility metric. This assumes that these sources of uncertainty are normally distributed. This is an acceptable assumption for the emulator and code uncertainty due to the Gaussian form of the predictions. However, the assumption should be checked for the observational and model discrepancy uncertainties for a given application.

The proposal distribution used in importance sampling must have adequate support. Here a multivariate Gaussian distribution is used,

$$q^{un}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mu_{nI}, \kappa\Sigma_{nI}) \tag{13}$$

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, due to the bounded approach for sampling the simulator in each wave. This

means the weights in Eq. (11) become $w^{un} = L(\boldsymbol{\theta}_q)/q^{un}(\boldsymbol{\theta}_q)$ where $\boldsymbol{\theta}_q$ are a number of samples from q^{un} and the constant prior essentially truncates the proposal samples to be within the final non-implausible domain.

Lastly the approximate posterior from Eq. (11) can be re-sampled in order generate direct samples from the approximate posterior. This involves drawing N_q samples where the probability of occurrence is defined by the normalised weights $w(\boldsymbol{\theta}_q) = w^{un}(\boldsymbol{\theta}_q)/\sum w^{un}(\boldsymbol{\theta}_q)$.

Fig. 2 demonstrates importance sampling and re-sampling applied to the numerical example in Fig. 1, where $N_q = 10,000$ and $\kappa = 2$. The re-sampled posterior samples are subsequently used to draw Monte Carlo realisations of the simulator and bias-corrected output. The results show that the emulator has been adequately calibrated with the two parameter solutions lying within the central probability mass. Furthermore the simulator and bias-corrected results lie within the given uncertainty bounds. These uncertainty bounds are $\pm 3\sqrt{V_o + V_m}$ for the simulator, reflecting both the model discrepancy and observational uncertainties, and $\pm 3\sqrt{V_o}$ for the bias-corrected output, where 3 standard deviations reflect Pukelsheim's 3σ rule.

3. Numerical case study: mass, tensioned wire system

BHM accounts for model discrepancy by defining a prior variance V_m, stating an assumption of uniform additive discrepancy across the space. In order to illustrate its effectiveness a simple numerical example of a mass, tensioned wire system, depicted in Fig. 3, is presented. In this case study the simulator output is the natural frequency of the system $y = \omega_n$, with the input being an applied tension $x = T$, and the calibration parameter being the mass $\theta = M$. The simulator is formed from,

$$\eta(x, \theta) = \omega_n(T, M) = \frac{1}{\pi}\sqrt{\frac{T}{Ml}} \tag{14}$$

where $l = 1\text{m}$ is the length. In this example model discrepancy is additive and sinusoidal defined as, $\delta(x) = 0.5\sin(2\pi \times 0.044x)$. This form of model discrepancy is chosen as it could be defined as uniform and additive, which is the assumption made within BHM. The observational data is obtained from,

$$z(x) = \eta(x, 5.43) + \delta(x) + e \tag{15}$$

where the 'true' mass is $\hat{\theta} = 5.43$ and the observational uncertainty is $e \sim \mathcal{N}(0, 0.01)$. Fig. 4a presents differences between the simulator output (using the 'true' mass) and the realisation observational responses (where $z(x_*)$ are 50 realisations of Eq. (15)).

The training data is shown in Fig. 4a, where the inputs are in 100 N steps from 200 N to 1000 N. The observational uncertainties are depicted on the training data as $\pm\sqrt{(V_o + V_m)}$ intervals. In this numerical example the variances associated with these uncertainties are known; $V_o = 0.01$ and $V_m = 1/12$ (calculated as the variance from a uniform distribution bounded [-0.5 0.5]). A simulator budget of 15 runs was given for defining the parameter space, as evaluations of the simulator were cheap for this case study; this number of runs allowed the emulator to adequately estimate the simulator in one wave, allowing the case study to focus on the methods ability to account for model discrepancy. As the calibration example is one-dimensional, i.e. only the mass is calibrated, a uniform grid is used to sample the space between [2 20] kg. The emulator utilised in this case study was constructed from a linear mean function $m(\mathrm{x}, \boldsymbol{\theta}) = [\mathrm{x}, \boldsymbol{\theta}]^{\mathrm{T}}\boldsymbol{\beta}$, reflecting the prior assumption that the natural frequency will increase with tension. In addition, a Matérn 3/2 covariance function was implemented, defining a prior that assumes the simulator is a relatively smooth function.

A multivariate implausibility metric was implemented with a threshold calculated from the 99% quantile of a 9 degree-of-freedom χ^2-distribution. 50,000 uniformly distributed samples were used to explore the parameter domain and the approximate posterior was sampled 10,000 times. Due to the emulator being trained from an adequate number of simulator runs, BHM reached the stopping criteria after one wave. The approximate posterior is depicted in Fig. 4b, where the 'true' mass (5.43 kg) is shown in red and the MAP estimate (5.44 kg) in blue. The mean squared error between the simulator (with the 'true' mass) and the MAP estimate from BHM for the test inputs was 2×10^{-5}Hz, showing an excellent fit; where the test inputs were 200 evenly space points between [200 1000]N.

This simple case study demonstrates the applicability of BHM for calibrating a simulator whilst accounting for model discrepancy. This well-controlled case study provides motivation for applying BHM to a more complex experimental case study in section 4.

4. Experimental case study: five storey shear structure

An experimental case study of a representative five storey building structure is presented in order to demonstrate the effectiveness of BHM. The aim of calibration in this scenario was to infer the material properties $\boldsymbol{\theta} = \{E, \nu, \rho\}$ (elastic modulus, Poisson's ratio and density respectively) of the columns and floors of an Finite Element (FE) model, using observational data from a representative building structure made from aluminium 6082. In this case study the simulator has been created such that model-form errors, caused by a simplification of the attachments to ground and of the bolted joints (modelled as bonded), are present. These issues cause a clear model discrepancy particularly in the first natural frequency, highlighting

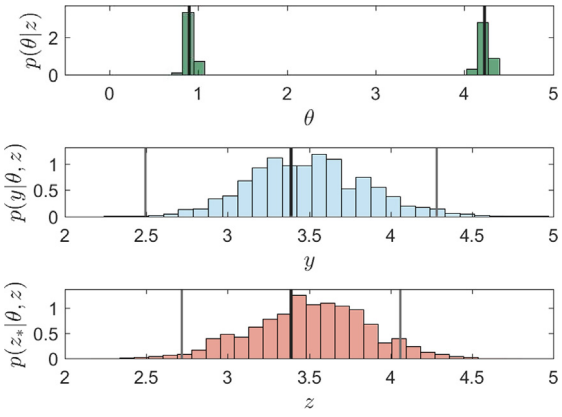

Fig. 2. Posterior and predictive samples from a BHM numerical example. The top panel shows the approximate posterior $p(\boldsymbol{\theta}|z)$. The middle panel presents the simulator output $p(y|\boldsymbol{\theta},z)$ given these posterior samples, where the black line denotes the 'true' value and the grey lines are $\pm 3\sqrt{V_o + V_m}$. The bottom panel shows the bias-corrected output $p(z_*|\boldsymbol{\theta},z)$ (where $z_* = z - e$) given the posterior samples, where the black line denotes the 'true' value and the grey lines are $\pm 3\sqrt{V_o}$.

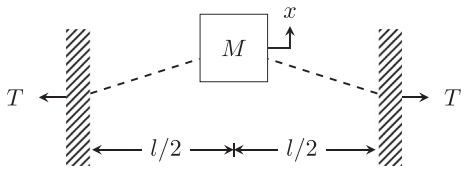

Fig. 3. Schematic of the mass, tensioned wire system.

the need for BHM—it is noted that despite these model-form errors, only the material properties are calibrated. The quantities of interest were the first five bending natural frequencies of the structure under different masses, $m = \{0, 0.1, \ldots, 0.5\}$ kg, added to the fourth floor of the building—simulating pseudo-damage extents.

The calibration process is presented in section 4.1, where the posterior distribution over the parameters is identified. Following calibration, section 4.2 outlines a method for inferring the functional form of the model discrepancy, leading to predictions that are bias-corrected, i.e. account for the model discrepancy. Validation of these bias-corrected predictions is subsequently performed in section 4.3, highlighting the benefits of the novel combined approach.

Modal testing was performed via an electrodynamic shaker; where the setup is shown in Fig. 5. The excitation was band-limited Gaussian noise, with a bandwidth of 409.6 Hz; where 40 averages were obtained for each test. The sample rate and time were chosen such that the frequency resolution was 0.05 Hz. Five accelerometers were placed on each of the five floors to measure the first five bending modes. For each mass, ten repeats were conducted in order to obtain an understanding of the underlying modal frequency distributions.

The simulator $\eta(\boldsymbol{x}, \boldsymbol{\theta})$ was a modal FE 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. The model parameters $\boldsymbol{\theta}$ were: elastic modulus E, Poisson's ratio ν and density ρ. The prior bounds on $\boldsymbol{\theta}$ were $\pm 15\%$ of typical material properties for aluminium 6082, shown in Table 1. A fifteen point, three dimension GMLHC was used to sample the parameter space, with an independent five point, three dimension GMLHC for validation.

The training observational data $z(\boldsymbol{x})$ used within the calibration process were the average natural frequencies when $\boldsymbol{x} = \{0, 0.3, 0.5\}$ kg. The unseen validation data set $z(\boldsymbol{x}_*)$ were the full repeat measurements at $\boldsymbol{x}_* = \{0.1, 0.2, 0.4\}$ kg.

4.1. Bayesian history matching

BHM was implemented as outlined in Algorithm 1. Five independent GP emulators were constructed from the training GMLHC, such that the output natural frequencies could be predicted[1]. Each of the five emulators were constructed from a linear mean function $m(\boldsymbol{x}, \boldsymbol{\theta}) = [\boldsymbol{x}, \boldsymbol{\theta}]^T \boldsymbol{\beta}$ and Matérn 3/2 covariance function. Exploration of the parameter domain was performed via propagating 100,000 samples from a uniform distribution over the bounds through the GP emulators (where each emulator was trained from fifteen simulator runs and validated against five separate simulator runs).

[1] It is noted that the simulator outputs will not be independent; an area of further research is in implementing multivariate GP emulators [25,26] within BHM.

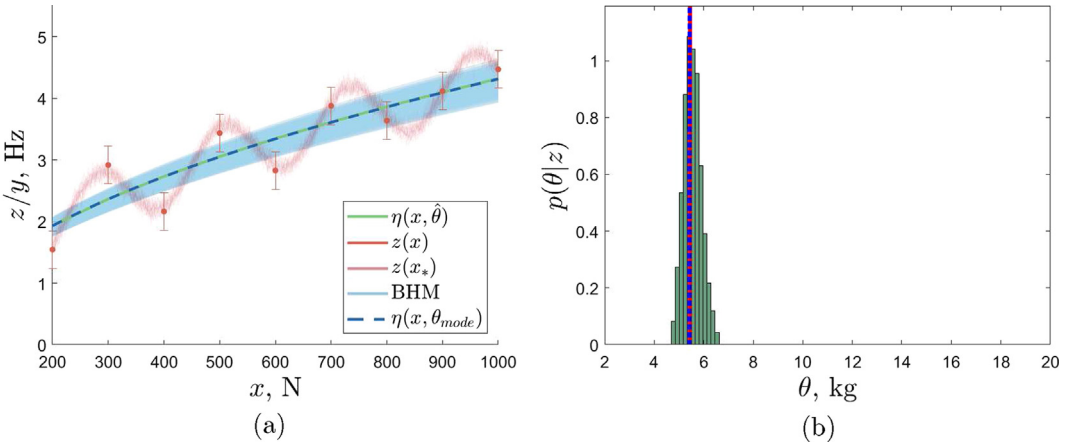

Fig. 4. A comparison of observational data, simulator predictions and BHM inferences. Panel (a) presents the output predictions, where the red points (•) are the observational data with $\pm\sqrt{V_o + V_m}$ bounds. Panel (b) shows the predicted posterior distribution with the MAP estimate (blue) against the true mass value (red).

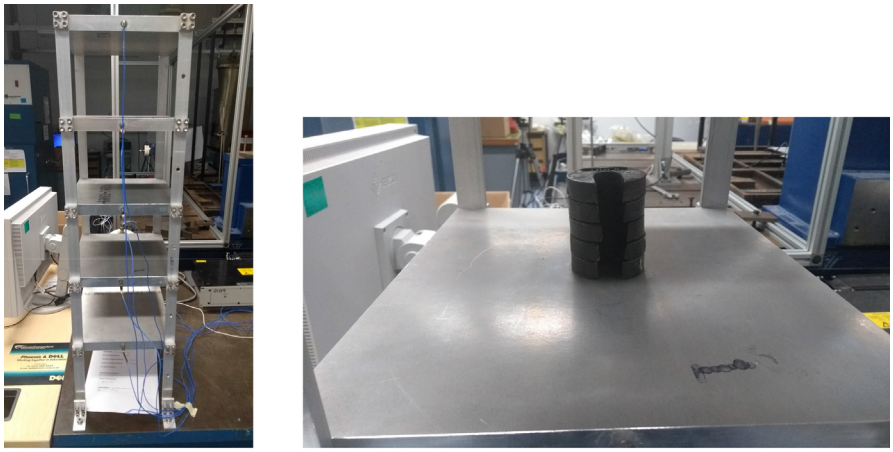

Fig. 5. Representative five storey building structure. Panel (a) show the test setup and panel (b) presents an example of the pseudo-damage, glued added masses, applied to the fourth floor.

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

Parameter		Lower Bound	Upper Bound
Elastic Modulus	E	60.35 GPa	81.65 GPa
Poisson's Ratio	ν	0.2805	0.3795
Density	ρ	2354.5 kg/m^3	3185.5 kg/m^3

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

Uncertainty		ω_1	ω_2	ω_3	ω_4	ω_5
Observational	V_o	0.003	0.001	0.01	0.331	0.141
Model Discrepancy	V_m	1.50	0.01	0.01	0.01	0.01

The multivariate implausibility metric (Eq. (7)) 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 the training inputs $\boldsymbol{x}$). Table 2 outlines the process uncertainties utilised in the implausibility metric. Observational variances $V_{o,j}$ were estimated for each natural frequency from the variance of the output training data. The model discrepancy variances were determined using expert judgement; the simulators prediction of the first natural frequency was know to provide poor predictive performance and therefore was given a large variance, whereas the other remaining outputs were more accurate, leading to smaller prior model discrepancy variances.

The stopping criteria was met after one wave, as the emulators inferred code uncertainties were less than the total process uncertainties for each output—diagnostic checks using methods outlined in [19] were also used to confirm the inferred emulators were valid. After the first wave a non-implausible space $\approx 2.8\%$ of the original space was identified.

In order to visualise the non-implausible space, minimum implausibility and optical depth plots were created. These quantities divide the parameter space into 'bins' within which each of the $100,000$ samples (from the uniform parameter domain sampling) are placed. Minimum implausibility takes the lowest value of implausibility below the threshold for the set of samples within a given bin. This provides an indication of which parts of the parameter space can be discarded irrespective of the other parameters. Optical depth is the ratio between non-implausible samples and the total number of samples within a given bin, providing an estimate of the probability of finding a non-implausible parameter combination given the set within a bin. Fig. 6 presents these quantities after the first wave when each parameter is divided into thirty bins. Here it can be seen that the outputs, as expected, are relatively insensitive to changes in Poisson's ratio. Furthermore, there is a clear linear correlation between the non-implausible space of the elastic modulus and density, displayed in the bottom left and top right quadrants of Fig. 6.

Once the stopping criteria has been met approximate posterior densities can be formed using importance sampling and re-sampling. In this case, a Gaussian proposal distribution with $\kappa = 1.5$ was used to generate $100,000$ samples with which to assess the normalised weights using the methodology presented in Section 2.6. $100,000$ samples were subsequently obtained by re-sampling the posterior distribution. Fig. 7 presents the marginal and pairwise joint posterior distributions, which are visually similar to the minimum implausibility and optical depths; with a linear relationship between density and elastic modulus and a relatively insensitive effect from Poisson's ratio in the pairwise joint distributions. Fig. 8 displays the marginal posterior distribution for each parameter, all showing slightly bi-modal distributions. The marginal distributions are in contact with the bounds of the prior parameter domain. This demonstrates a known strength of BHM, as the bounds constrain calibration to physical values. If these bounds were not used then non-physical parameter inferences may be arrived at, due to the existence of model discrepancy.

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 were low, each emulator mean was taken as deterministic and the GPs were not sampled. The mean predictions of the GP emulators for the 1000 Monte Carlo realisations are presented in Fig. 9. The predictions are shown against the observational data used within BHM $z(\boldsymbol{x})$ with $\pm c_\sigma \sqrt{V_{o,j} + V_{m,j}}$ bounds; where c_σ is the standard deviation associated with 99% probability mass of a standard normal (assuming output distributions to be approximately Gaussian). Fig. 9 demonstrates that all five outputs are within the defined uncertainty bounds. However large discrepancies between the experimental observations and simulator outputs (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

Inferring the functional form of the model discrepancy is important, as it provides a mechanism for targeting model improvements and quantifying simulator adequacy, i.e. a large model discrepancy will infer that the simulator may not be fit for purpose. In light of this aim, a method is proposed for inferring the functional form of the model discrepancy using the inferred parameter estimates from BHM. The approach involves utilising GP regression to infer the functional difference between the calibrated simulator output and the observational data, at a set of training points. This sections outlines and demonstrates the approach for the five storey building structure case study.

The method begins by obtaining the parameter MAP estimates from the posterior distribution identified by BHM, i.e. $\boldsymbol{\theta}_{MAP} = \max\, p(\boldsymbol{\theta}|z)$. The calibrated simulator output is subsequently obtained by evaluating the GP emulators at this calibrated value i.e. $p(\boldsymbol{\eta}_{*,j}|\boldsymbol{x}_*, \boldsymbol{\theta}_*, y_j, \boldsymbol{x}, \boldsymbol{\theta}_{MAP}, \boldsymbol{\phi})$, as the emulators have been established to approximate the simulator well during BHM. The mean prediction from these emulators, Eq. (4b), is taken to be the calibrated simulator output, $\eta(\boldsymbol{x}, \boldsymbol{\theta}_{MAP})$. Model discrepancy is subsequently modelled as a GP regression model (with a Gaussian noise variance) that maps between the calibrated simulator output (for all outputs) to the experimental data, i.e. $\mathcal{GP}_{j,\delta} : \eta(\boldsymbol{x}, \boldsymbol{\theta}_{MAP}) \rightarrow z_j(\boldsymbol{x})$. Once constructed, bias-corrected outputs are obtained $p(\boldsymbol{z}_{*,j}|\boldsymbol{x}_*, \boldsymbol{z}_j, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}_{MAP}, \boldsymbol{\phi}_{\delta,j})$, that account for model discrepancy. It is noted that this method does not propagate the parameter uncertainty from the posterior distribution through to the model discrepancy inference stage, and therefore this uncertainty is not reflected in the bias-corrected predictions. Uncertainty propagation of the param-

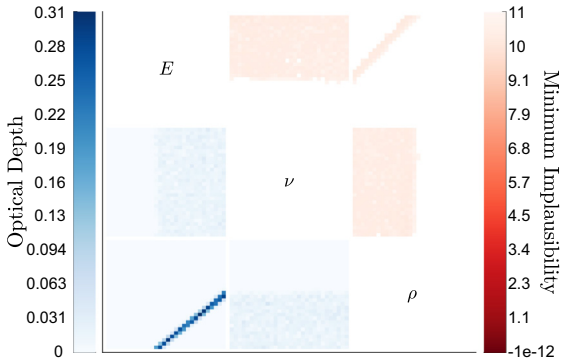

Fig. 6. Minimum implausibility and optical depth plots for the first wave of BHM on the representatives five storey building structure. Each quadrant is a comparison of two parameter combinations for the given metric, e.g. the top right quadrant is ρ against E for minimum implausibility and the bottom left E against ρ for optical depth.

eter uncertainty is therefore highlighted as an area of further research. Likewise, implementing the approach with multi-output GP regression models should also be pursued as further research.

The outlined approach was applied to the five storey building structure case study. Five independent GP regression models were again constructed but here mapping from the calibrated emulator outputs to the ten repeat observations (aiding estimation of the observational uncertainty) at the training inputs $\boldsymbol{x} = \{0, 0.3, 0.5\}$ kg. The GP priors were modelled using zero mean functions and Matérn 3/2 covariance functions. The bias-corrected output predictions are displayed in Fig. 10 and the inferred model discrepancy functions in Fig. 11.

The inferred model discrepancies in Fig. 11 capture part of the missing physics. It can be observed for the first, second and fourth natural frequencies that the discrepancies increase with mass, and that the first natural frequency has a large discrepancy of around 2 Hz. In contrast, the model discrepancies for the third and fifth natural frequencies indicate that the calibrated simulator closely matched the observational data, and therefore the model discrepancy functions likely capture a 'noise' process rather than any particular missing physics. These discrepancies would lead the modeller in targeting improvements to the simulator that corrects the first natural frequency most, and that account for the relatively linear increase in natural frequency with mass that the simulator currently fails to capture.

4.3. Validation

Validation metrics are applied to the bias-corrected predictions from the joint BHM and GP regression approach in order to assess the methods effectiveness. Normalised Mean Squared Errors (NMSEs) were used to assess the mean predictive performance; defined as,

$$NMSE = \frac{100}{N\sigma_{z*}^2} \sum \left(z_*(\boldsymbol{x}) - \hat{z}_*(\boldsymbol{x})\right)^2 \tag{16}$$

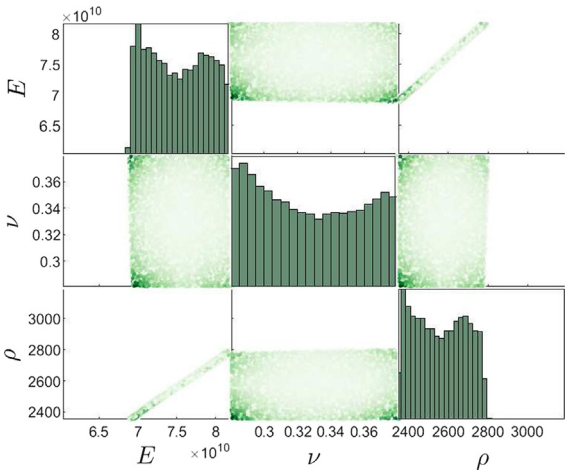

Fig. 7. Marginal and pairwise joint posterior distributions for the first wave of BHM on the representatives five storey building structure, where a darker shade represents a higher probability.

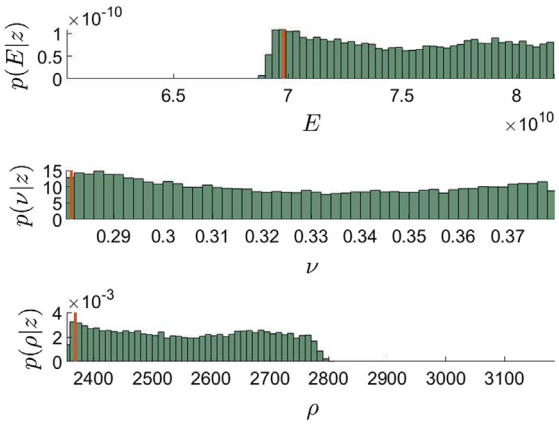

Fig. 8. Marginal posterior distributions for the first wave of BHM on a representatives five storey building structure. The red lines indicate the MAP estimates.

where $z_*(\boldsymbol{x})$ are the output test data, $\sigma^2_{z_*}$, the variance of the output test data and $\hat{z}_*(\boldsymbol{x})$, the mean test predictions. A score of zero indicates a mean prediction without any error; conversely, a score of 100 represents a scenario where the prediction is no better than taking the mean of the true values $\mathbb{E}(z_*(\boldsymbol{x}))$. Table 3 presents the scores for each natural frequency demonstrating good predictive performance (as they are all below five). Predictive performance was found to be poorest for the second natural frequency, which is due to the discrepancy when $x = 0.1$ kg.

The approach in this paper is probabilistic and as such the complete predictive distributions should be assessed. In this case, three statistical distances are applied: the Area Metric, Hellinger distance and Maximum Mean Discrepancy (MMD) distance, which measure the distance between two probability measures, $\mathbb{P}$ and $\mathbb{Q}$. The Area Metric [27] is the L_1-norm between two CDFs ($F(x)$),

$$D_{Area}(\mathbb{P},\mathbb{Q}) = \int |F_P(x) - F_Q(x)| \mathrm{d}x \tag{17}$$

where the units of the Area Metric are that of the quantities of interest. The units therefore make this a useful validation metric. Additionally, the Area Metric assess the distance between CDFs, this means that an empirical CDF can be used to provide a non-parametric estimation of the observation distributions. The second validation metric considered is the Hellinger distance; the L_2-norm between two Probability Density Functions (PDFs) ($p(x)$),

$$D_H(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{2}\int \left(\sqrt{p_P(x)} - \sqrt{p_Q(x)}\right)^2 \mathrm{d}x}. \tag{18}$$

where the metric is bounded [0 1]. These bounds mean the Hellinger distance is good at objectively comparing predictions at different scales, as zero means the distributions are the same, and one, that the distributions are completely different. The final validation metric is the MMD distance, which is the maximum distance between the mean embeddings of two sample sets in a RKHS, calculated as,

$$D_{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_x(f(x)) - \mathbb{E}_y(f(y))\right| \tag{19}$$

where the projection is performed by the function class $\mathcal{F}$, where the function f is called a reproducing kernel $k(\cdot,\cdot)$ [28], and x and y are samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively. The distance is non-parametric and has a lower bound of zero. A popular choice of kernel is the radial basis kernel [28,29], where the median pairwise distance among the joint data is commonly used to infer the hyperparameters [30].

The three validation metrics are applied to the output predictions and observational samples for each of the five natural frequency. Numerical integration is used to infer both the Area Metric and Hellinger distance, where an empirical CDF and kernel density estimate are used to approximate the observational distribution for both the Area Metric and Hellinger distance respectively. Due to the MMD being sample based, ten samples were obtained from predictive distributions such that the distance could calculated between these samples and the observational samples. This procedure was repeated 100 times in order to obtain the average MMD distance, which should be more robust to the predictive distribution sampling. Furthermore, the MMD distance was implemented using a radial basis kernel, reflecting the expected Gaussian-form of the observational data, where the hyperparameters were inferred using the median heuristic. The distances are presented in Fig. 12.

The Area Metric values, shown in Fig. 12a, are relatively small—of the order 10^{-3}—providing evidence that the predictive distributions are close the observational data. The largest distance is for the fifth natural frequency at 0 kg, which is likely due to the spread of observation samples at 0 kg, with one data point, which potentially is an outlier, lying far from the

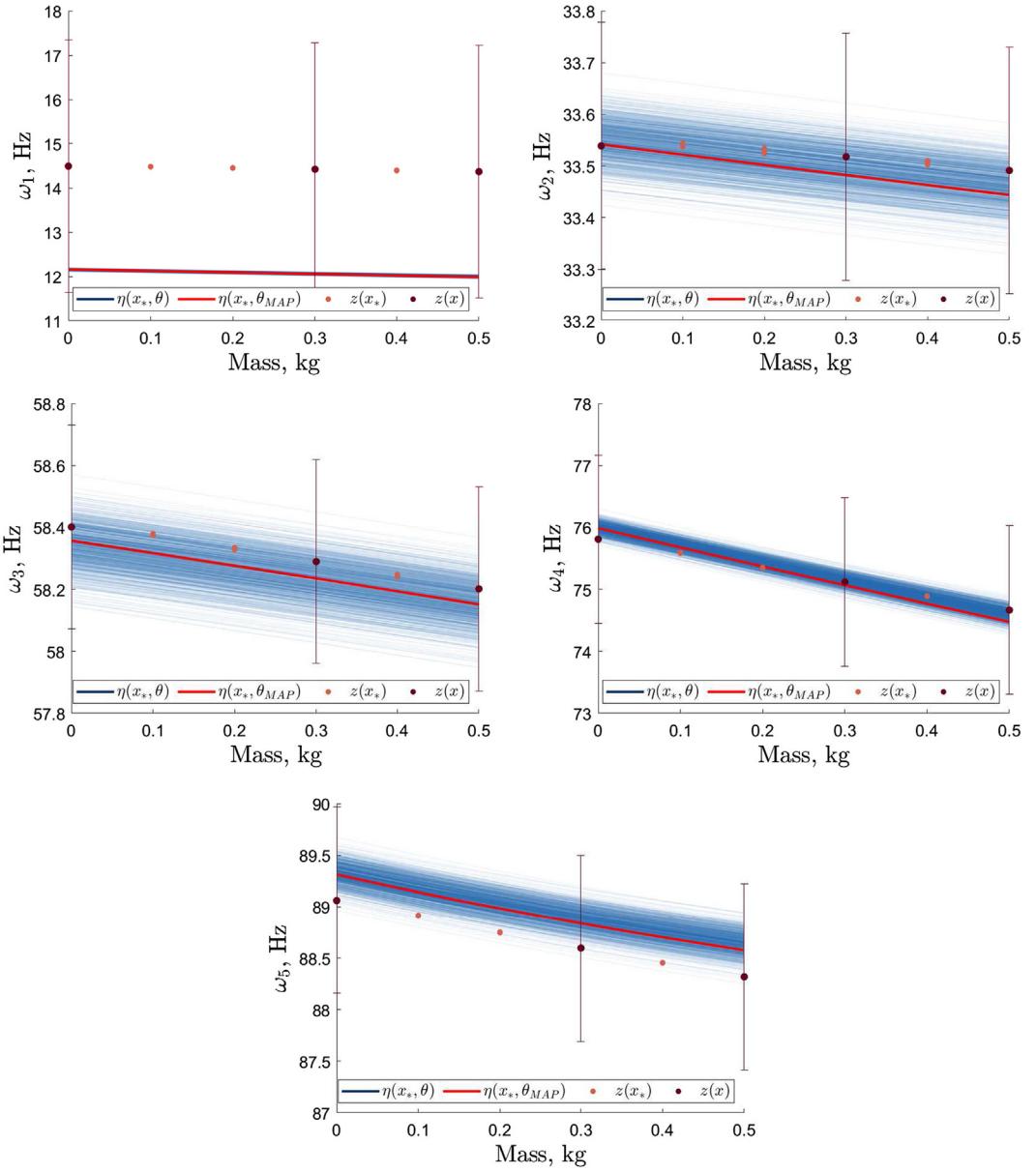

Fig. 9. 1000 samples of the BHM predictive outputs, $\mathbb{E}(p(\boldsymbol{\eta}_{*j}^{(i)}|\boldsymbol{x}_*, \boldsymbol{y}_j, \boldsymbol{x}, \boldsymbol{\theta}^{(i)}, \boldsymbol{\phi}_{\eta j}))$ given $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|Z, \boldsymbol{x}_z)$. The error bounds are $\pm c_\sigma \sqrt{V_o + V_m}$.

others. The next relatively large distances are for the 0.1 kg case for all natural frequencies. These are due to the offset in the predictive mean. This result is further evidenced in the Hellinger distances, Fig. 12b, where the 0.1 kg case produces relatively large distances, i.e. close to 0.5. The MMD distances also identify the 0.1 kg case as producing the largest distances. This discrepancy at the 0.1 kg mass input is due to the training data not providing enough information about the functional trend for this input. More training data would therefore improve the predictive performance at this location. Additionally, improvements to the simulator could aid predictions at 0.1 kg.

Finally, the MMD witness function was evaluated for the second natural frequency in order to further investigate its performance, due to it having the largest NMSE. The MMD witness function $f^*(\cdot)$, is the difference between two kernel embeddings, defined over the variable t as,

$$f^*(t) \propto \frac{1}{m}\sum_{i=1}^{m} k(x_i, t) - \frac{1}{n}\sum_{j=1}^{n} k(y_i, t) \tag{20}$$

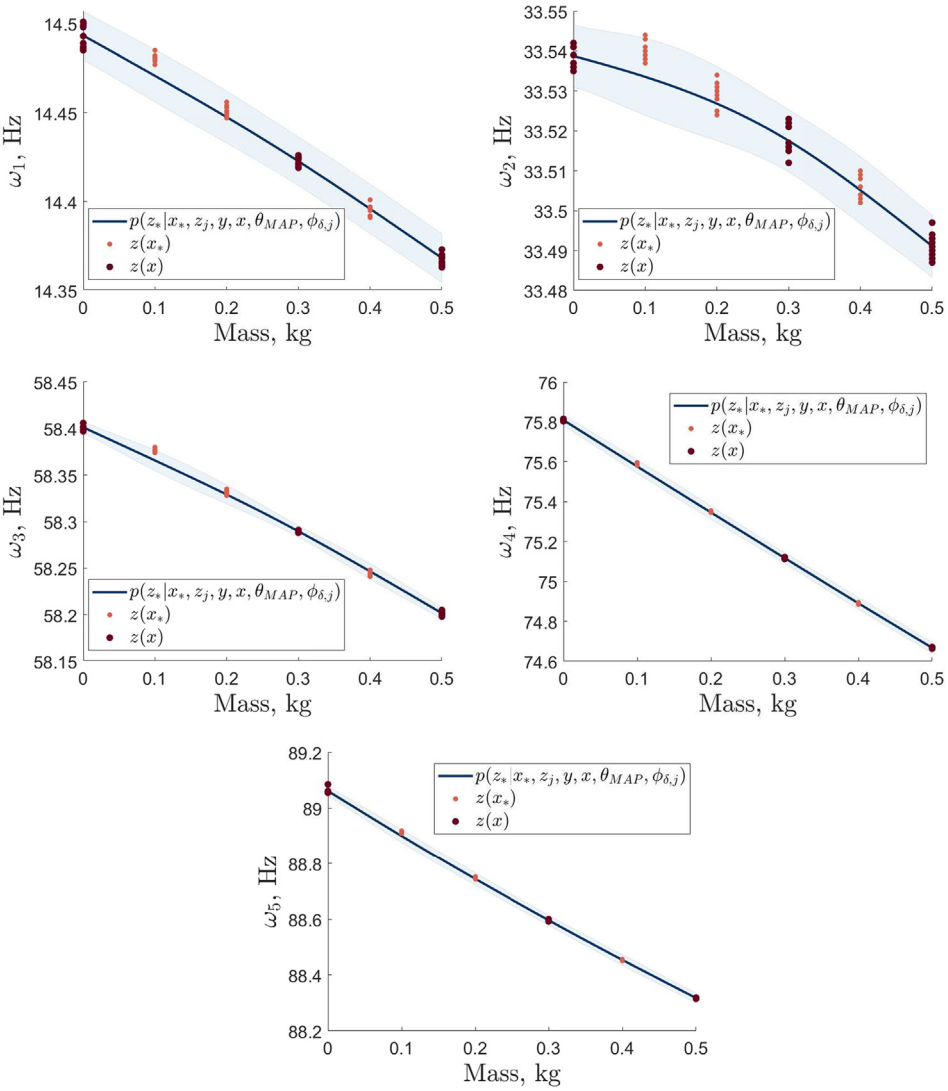

Fig. 10. The bias-corrected predictive outputs from the combined BHM and GP regression approach; the shaded regions indicate $\pm 3\sigma$.

where $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ are the two sample sets. The witness function provides a visualisation of the difference between two distributions and is zero intuitively where the two distributions are the same, positive when $\mathbb{P}$ is larger than $\mathbb{Q}$, and negative when $\mathbb{Q}$ is greater than $\mathbb{P}$, as far as the smoothness constraint allows. Therefore, high values, whether positive or negative, indicate a large difference between the predictive and observational distributions.

Fig. 13 presents the witness function for the second natural frequency predictions. The 0, 0.3, 0.4 and 0.5 kg cases have low witness function magnitudes, which is expected for the training data $\boldsymbol{x} = \{0, 0.3, 0.5\}$ kg but shows good predictive performance for 0.4 kg. It can also be seen that the 0.4 kg predictive distribution is narrower than the observational data (indicated by negative values about the mode), meaning the results are not conservative. The mode is also under-estimated for the 0.1 and 0.2 kg cases however, the variance for the 0.2 kg case is conservative, covering the observational prediction. Inclusion of the parameter uncertainty may inflate these predictive distributions, potentially improving predictive performance, or at least making the distributions more conservative.

5. Conclusions

Model discrepancy poses challenges in calibrating structural dynamics simulators. Without accounting for the presence of model discrepancy, any parameter estimate will be biased and predictive performance potentially poor. In this paper it has been demonstrated that BHM provides a methodology for calibrating simulators whilst assuming an additive model discrepancy. The approach has been demonstrated to be successful on both a numerical case study and on an experimental five

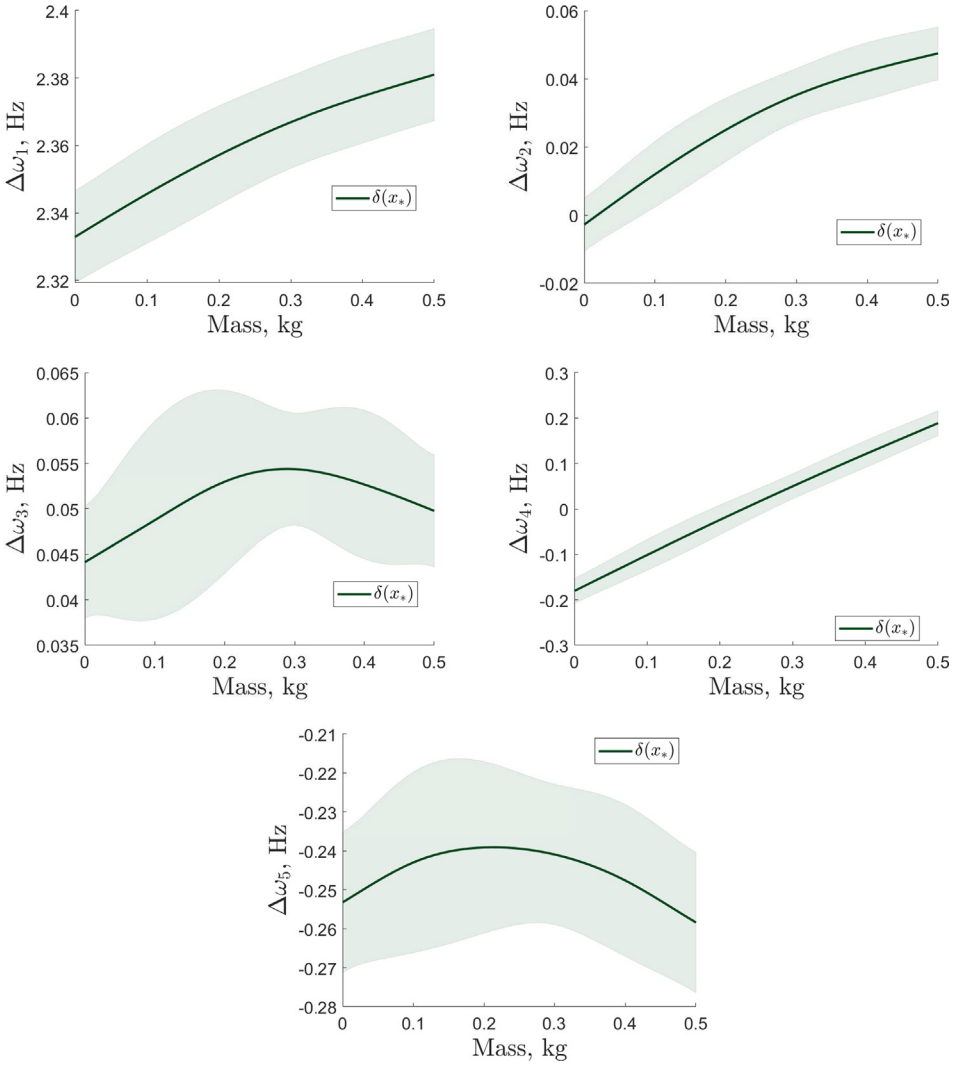

Fig. 11. Inferred model discrepancies; the shaded regions indicate $\pm 3\sigma$.

Table 3
Normalised mean squared errors (NMSEs) between bias-corrected output predictions and experimental data.

	ω_1	ω_2	ω_3	ω_4	ω_5
NMSE	1.51	4.31	0.47	0.02	0.08

storey building structure. Furthermore, a method has been outlined for inferring model discrepancy functional forms. The novel combined technique has been shown to provide improved predictive performance and a greater insight into simulator inadequacies.

BHM is an effective method for discarding parameter space iteratively in a 'likelihood-free' manner. This approach means that difficult-to-emulate outputs, or input combinations, can be excluded and reintroduced between waves when they are more defined; this is not possible in a likelihood based approach. By separating parameter inference from model discrepancy learning the technique removes non-identifiability problems, provided its assumptions are appropriate. Furthermore, by utilising importance sampling approximations of the posterior parameter distribution can be obtained. There are avenues for further research into optimal sampling methods for assessing the parameter space, with sequential design methods potentially providing an effective method for reducing the number of simulator evaluations required to construct effective emulators. Multivariate GP emulators should also be incorporated, such that a more informative prior can be used for dependent outputs.

The model discrepancy inference technique outlined in this paper, constructs GP regression models that map the calibrated simulator outputs to experimental observations. This approach has been demonstrated to be effective in learning such func-

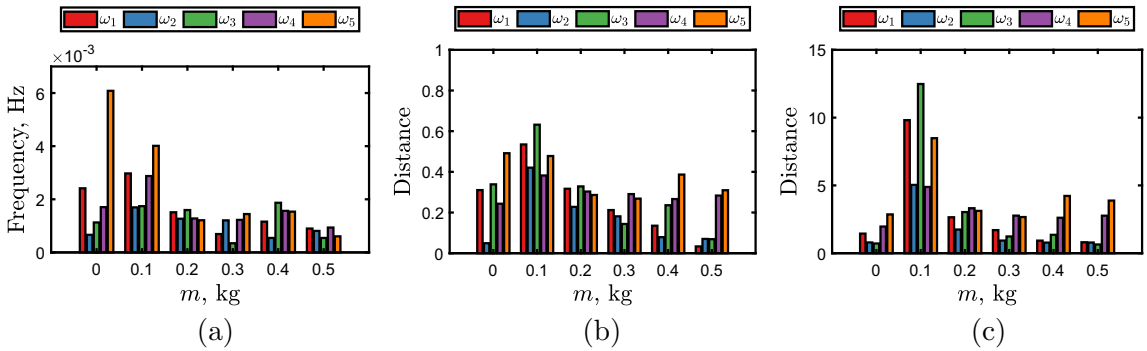

Fig. 12. Statistical distances between bias-corrected output predictions and experimental data. Panel (a) shows the Area Metric, panel (b), the Hellinger distance and panel (c) the MMD distance.

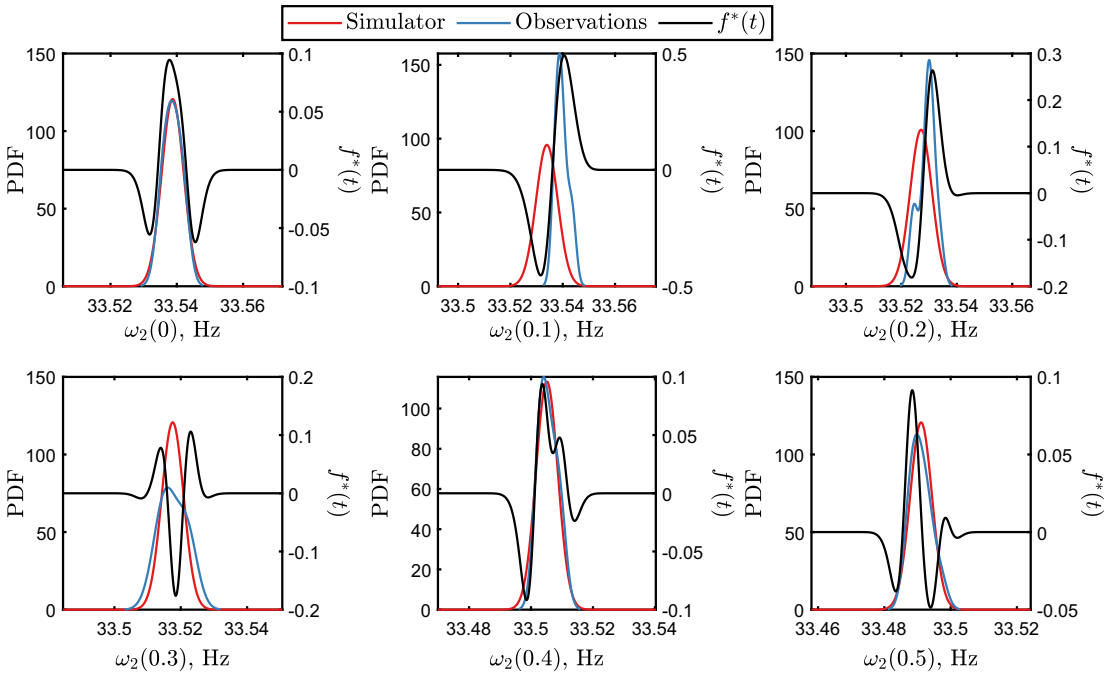

Fig. 13. Witness function between the bias-corrected output predictions and experimental data for the second natural frequency.

tions from a small number of data points. Further research should be conducted into propagating the full parameter distributions in a computationally efficient manner, without the need for constructing a large quantity of GP models. Additionally, research should be conducted into physically constraining these GPs such that prior knowledge is utilised effectively. Nonetheless, the work presented in this paper demonstrates an effective methodology for both calibrating computer models when model discrepancy is present, and inferring the functional form of that model discrepancy. The approach has been shown to improve predictive performance and aid in identification of improvements to the computer model.

CRediT authorship contribution statement

P. Gardner: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Software. **C. Lord:** Writing - review & editing, Supervision. **R.J. Barthorpe:** Conceptualization, Funding acquisition, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) via Grant Nos., EP/R006768/1 and EP/N010884/1.

References

[1] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc.: Ser. B 63 (3) (2001) 425–464.
[2] J. Brynjarsdóttir, A. O'Hagan, Learning about physical parameters: the importance of model discrepancy, Inverse Prob. 30 (11) (2014) 114007.
[3] D.S. Oliver, Y. Chen, Recent progress on reservoir history matching: a review, Comput. Geosci. 15 (1) (2011) 185–221.
[4] M.I. Friswell, J.E. Mottershead, Inverse methods in structural health monitoring, Key Eng. Mater. 204–205 (2001) 201–210.
[5] M.I. Friswell, J.E. Mottershead, Physical understanding of structures by model updating, in: Proceedings of International Conference on Structural System Identification, 2001, pp. 81–96.
[6] P.S. Craig, M. Goldstein, A.H. Seheult, J.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, 1997, pp. 37–93.
[7] I. Vernon, M. Goldstein, R.G. Bower, Galaxy formation: a Bayesian uncertainty analysis, Bayesian Anal. 5 (4) (2010) 619–669.
[8] M. Goldstein, External Bayesian analysis for computer simulators, in: Bayesian Statistics 9, number 1996, Oxford University Press, 2011, pp. 201–228.
[9] I. Vernon, M. Goldstein, R. Bower, Galaxy formation: Bayesian history matching for the observable universe, Stat. Sci. 29 (1) (2014) 81–90.
[10] I. Andrianakis, I.R. Vernon, N. McCreesh, T.J. McKinley, J.E. Oakley, R.N. Nsubuga, M. Goldstein, R.G. White, Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda, PLoS Comput. Biol. 11 (1) (2015) e1003968.
[11] I. Andrianakis, I. Vernon, N. McCreesh, T.J. McKinley, J.E. Oakley, R.N. Nsubuga, M. Goldstein, R.G. White, History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation, J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 66 (4) (2017) 717–740.
[12] N.R. Edwards, D. Cameron, J. Rougier, Precalibrating an intermediate complexity climate model, Clim. Dyn. 37 (7–8) (2011) 1469–1482.
[13] D. Williamson, M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, K. Yamazaki, History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Clim. Dyn. 41 (7–8) (2013) 1703–1729.
[14] D. Higdon, J. Gattiker, B. Williams, M. Rightley, Computer model calibration using high-dimensional output, J. Am. Stat. Assoc. 103 (482) (2008) 570–583.
[15] P.D. Arendt, D.W. Apley, W. Chen, Quantification of model uncertainty: calibration, model discrepancy, and identifiability, J. Mech. Des. 134 (10) (2012) 100908.
[16] P.D. Arendt, D.W. Apley, W. Chen, A preposterior analysis to predict identifiability in the experimental calibration of computer models, IIE Trans. 48 (1) (2016) 75–88.
[17] A. O'Hagan, J.F.C. Kingman, Curve fitting and optimal design for prediction, J. R. Stat. Soc. Ser. B 40(1) (1978) 1–42.
[18] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[19] L.S. Bastos, A. O'Hagan, Diagnostics for Gaussian process emulators, Technometrics 51 (4) (2009) 425–438.
[20] I. Andrianakis, P.G. Challenor, The effect of the nugget on Gaussian process emulators of computer models, Comput. Stat. Data Anal. 56 (12) (2012) 4215–4228.
[21] P. Gardner, T.J. Rogers, C. Lord, R.J. Barthorpe, Sparse gaussian process emulators for surrogate design modelling, Appl. Mech. Mater. 885 (2018) 18–31.
[22] K. Worden, G. Manson, N.R.J. Fieller, Damage detection using outlier analysis, J. Sound Vib. 229 (3) (2000) 647–667.
[23] F. Pukelsheim, The three sigma rule, Am. Stat. 48 (2) (1994) 88–91.
[24] H. Dette, A. Pepelyshev, Generalized latin hypercube design for computer experiments, Technometrics 52 (4) (2010) 421–429.
[25] P. Boyle, M. Frean, Dependent Gaussian processes, in: Advances in Neural Information Processing Systems, 2005, pp. 217–224.
[26] T.E. Fricker, J.E. Oakley, N.M. Urban, Multivariate Gaussian process emulators with nonseparable covariance structures, Technometrics 55 (1) (2013) 47–56.
[27] S. Ferson, W.L. Oberkampf, L. Ginzburg, Model validation and predictive capability for the thermal challenge problem, Comput. Methods Appl. Mech. Eng. 197 (29–32) (2008) 2408–2430.
[28] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-Sample test, J. Mach. Learn. Res. 13 (1) (2012) 723–773.
[29] J.R. Lloyd, Z. Ghahramani, Statistical model criticism using kernel two sample tests, in: Advances in Neural Information Processing Systems, 2015, pp. 829–837.
[30] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, A.J. Smola, A kernel statistical test of independence, in: Neural Information Processing Systems, 2008, pp. 585–592.