Proceedings of the ASME 2019
Verification and Validation Symposium
VVS2019
May 15-17, 2019, Las Vegas, NV, USA

VVS2019-5149

SEQUENTIAL BAYESIAN HISTORY MATCHING FOR MODEL CALIBRATION

Paul Gardner*, Charles Lord, and
Robert J. Barthorpe
Dynamics Research Group
Department of Mechanical Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
Email: p.gardner@sheffield.ac.uk

ABSTRACT

Computer models, whilst frequently utilised for many complex engineering tasks, suffer from model form errors due to some level of simplification and/or absence of certain physics. These model form errors lead to a mismatch between model outputs and observational data when the 'true' parameters are known; a phenomenon known as model discrepancy. Calibration of a computer model without consideration of this type of uncertainty therefore leads to biased estimates of system parameters. Bayesian history matching (BHM) is one such method of calibrating a computer model whilst accounting for uncertainties associated with model discrepancy. The 'likelihood-free' technique assesses the system parameter domain using an emulator of the complex computer model in order to discard parameter combinations based on how unlikely they were to have produced a known observation response. BHM can be approached in an iterative manner, allowing sequential-based approaches to be used in selecting new computer model evaluations that will maximise the improvement in emulator performance. This paper develops techniques for sequentially selecting new computer model evaluations, reducing the total number of evaluations and increasing improvements in the emulator. The developed metrics and criteria are outlined with a demonstration on a numerical case study in order to visually demonstrate their applicability and increase in computational efficiency.

*Address all correspondence to this author.

INTRODUCTION

Model form errors, inherent to all models due to the absence and/or simplification of certain physics, cause a mismatch between observational data and computer model (or *simulator*) outputs when given the 'true' parameters. This form of uncertainty, known as model discrepancy, must be accounted for within any calibration process, otherwise the inferred parameter estimates will be biased. Bayesian History Matching (BHM) is a methodology for calibrating simulators under the assumption of model discrepancy. The 'likelihood-free' approach creates a rejection criteria based on a measure of dissimilarity between input-parameter combination outputs and known observational data, given various uncertainty sources including model discrepancy. As a result input-parameter pairs can be removed and added back into the analysis without invalidating the procedure and that the domain can be truncated based on physical understanding of the problem, reducing non-identifiablity issues.

BHM was first developed for the oil industry [1] but since 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]. The methods implemented within the literature rely on an iterative procedure; parts of the input-parameter domain are discarded and then an emulator (also known as a surrogate model) is retrained using new simulator evaluations — such that it more accurately reflects the output response surface in non-implausible regions. However, during each iteration these new simulator evaluations are determined by space-filling de-

signs. This paper therefore investigates sequential design procedures for selecting new evaluations locations, such that fewer simulator runs are required, improving the computational efficiency of BHM.

The outline of this paper is as follows; an overview of BHM is presented followed by the presentation of a numerical case study. Sequential approaches are then discussed, with probability of non-implausible and expected (un)improvement being defined. Each approach is investigated on the numerical case study providing a discussion about their advantages when compared to a conventional space-filled methodology. Finally conclusions and future research are provided.

BAYESIAN HISTORY MATCHING

BHM seeks to calibrate a statistical model of the form shown in eq. (1) (in a similar formulation to that proposed in [8]).

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

Where $z_j(x)$ is the jth observational output given inputs x, $\eta_j(x,\theta)$ is the jth simulator given x and parameters θ. The model discrepancy and observational uncertainty are δ and e respectively. Equation (1) assumes that the simulator, model discrepancy and observational uncertainty are independent and does not seek to define the model discrepancy's functional form (although this may be useful in certain applications).

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 parameter combinations and discarded based on a metric and threshold. This process would be prohibitively computationally expensive in most applications if simulator runs were required for each proposed parameter combination. To reduce this computational burden an emulator is implemented, with common techniques being Gaussian Process (GP)s [4] and Bayes linear [2, 3] emulators. Here GP emulators are utilised, where the emulator is constructed as in eq. (2).

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

The predictive GP emulator mean $\mathbb{E}\left(\mathcal{GP}_j(x,\theta)\right)$ allows efficient assessment and exploration of the parameter space whilst also quantifying code uncertainty, $V_c(x,\theta) = \mathbb{V}\left(\mathcal{GP}_j(x,\theta)\right)$. The formulation stated in eq. (2) assumes univariate GP emulators for each output, however it is trivial to replace these with a multivariate GP formulation [9].

BHM employs a quantity that assesses the dissimilarities between observations and simulator outputs. A common metric is implausibility, which is the distance between observations and simulator outputs, weighted by the process's uncertainties, defined in eq. (3).

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

Where, V_o, V_m and $V_c(x,\theta)$ are the variances associated with the observational, model discrepancy and code uncertainties. By including code uncertainty $V_c(x,\theta)$ into eq. (3) parameter space is retained if the emulator variance is high for a particular parameter combination, meaning that space is not discarded until the emulator is more certain that it accurately represents the simulator in that region. 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 should be elicited from expert judgement; sensitivity analysis can be performed during a wave to understand changes in rejection rates. Observational and model discrepancy uncertainties can be dependant on both inputs x and outputs $z_j(x)$, i.e. $V_{o,j}(x)$ and $V_{m,j}(x)$, if input dependent hetroscedastic noise or model discrepancy are hypothesised.

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. Firstly, a maximum implausibility can be formed, whereby the worst case for a given parameter combination is used, defined in eq. (4).

$$I_{max}(\theta) = \arg\max_j \left(\arg\max_{x_i} I_j(x,\theta) \right) \tag{4}$$

The other approach is to form a multivariate implausibility metric for either the inputs or outputs, eqs. (5) and (6). This is equivalent to taking the mahalanobis distance, standard practice in outlier analysis [10], which assesses the euclidean distance of the principle components. Again a maximum can be taken over either eqs. (5) and (6) to collapse the metric to a single value for each parameter combination.

$$I_{multi}(\theta)_j = \left(z_j(x) - \mathbb{E}\left(\mathcal{GP}_j(x,\theta)\right)\right)^{\mathsf{T}} \left(V_{o,j} + V_{m,j} + V_{c,j}(x,\theta)\right)^{-1} \left(z_j(x) - \mathbb{E}\left(\mathcal{GP}_j(x,\theta)\right)\right) \tag{5}$$

$$I_{multi}(x,\theta) = \left(z_j(x) - \mathbb{E}\left(\mathcal{GP}_j(x,\theta)\right)\right)^{\mathsf{T}} \left(V_{o,j} + V_{m,j} + V_{c,j}(x,\theta)\right)^{-1} \left(z_j(x) - \mathbb{E}\left(\mathcal{GP}_j(x,\theta)\right)\right) \quad (6)$$

In order to decide which parts of the parameter space to exclude a decision should be made based on the implausibility metric, often taking the form of a threshold T. 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. A rejection criteria can be formed for a particular parameter combination θ as in eq. (7).

$$I(\theta) \begin{cases} \leq T & \text{if } \theta \in \theta_{nI}, \\ > T & \text{if } \theta \in \theta_I \end{cases} \quad (7)$$

The threshold value depends on the type of implausibility metric being considered. Andrianakis et al. state that a sensible threshold T for single $I_j(x,\theta)$ or maximum $I_{max}(\theta)$ implausibilities (where the maximum is of a single implausibility set) can be determined by Pukelsheim's 3σ rule [4]. 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 [11]. For multivariate implausibilities the threshold T can be set as a high percentile ($\alpha > 95\%$) from a chi-squared distribution with either j, or the input size of x, degrees of freedom [4], i.e. $T = F^{-1}_{\chi^2}(\alpha)$ 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.

Furthermore, the algorithm requires a method for sampling the parameter space in order to assess the criteria. A simple approach is to draw samples 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 for the standard approach a Generalised Maximum Latin Hypercube Design (GMLHD) based method is implemented [12]. An emulator is constructed from the simulator runs and its output assessed at parameter combinations sampled from a uniform distribution where the bounds are from the parameter domain. A set of these sample parameters can then be rejected based on the given metric and criteria, and the bounds of the non-implausible region determined. A new wave can then be run with a LHD constructed from the new 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. $V_{c,j}(x,\theta_{nI}) < V_{o,j} + V_{m,j}$, 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. Once identified the final non-implausible space can be used to perform approximate posterior sampling using an importance sampling technique [4].

Algorithm 1 Bayesian History Matching for Wave k

```
θ^k ~ GMLHC                                   ▷ Draw parameters from GMLHC
y^k = η(x, θ^k)                               ▷ Run the simulator at parameters
Draw n samples θ_s^k ~ U(min(θ^k), max(θ^k))
for j = 1 : no. of outputs do
    Train and validate GP_j(x, θ^k)           ▷ Train/validate emulators
    [E(GP_j(x, θ_s^k)), V_{c,j}(x, θ_s^k)] = GP_j(x, θ_s^k)   ▷ Predictions
    Calculate I_j(x, θ_s^k)                   ▷ Assess implausibility of samples
end for
Calcualate I_max(θ_s^k)
for m = 1 : n do
    if I_max(θ_{s,m}^k) < T then
        θ_{nI}^k = θ_{s,m}^k                  ▷ Keep non-implausible samples
    end if
end for
bounds = [min(θ_{nI}^k), max(θ_{nI}^k)]       ▷ Obtain new GMLHC bounds
if any (V_{c,j}^k(x, θ) < (V_{o,j} + V_{m,j})) or isempty(θ_{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. (8) models the experimental observation z, which is obtained from the 'true' process with noise, stated in eq. (9); 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 $\mathbb{V}((z-e)-y)$).

$$y = \eta(\theta) = 5.5\left(0.15\cos(2\pi \times 0.75\theta) + 1.25\sin(2\pi \times 0.1\theta)\right) \quad (8)$$

$$z(\theta) = y(\theta) - 0.3\sin(2\pi \times 0.15\theta) + e \quad (9)$$

Figure 1 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 $\theta_s = \{-0.5, 0.055, \ldots, 5\}$ where a budget of four simulator evaluations have been performed in a space-filling manner $\theta^1 =$

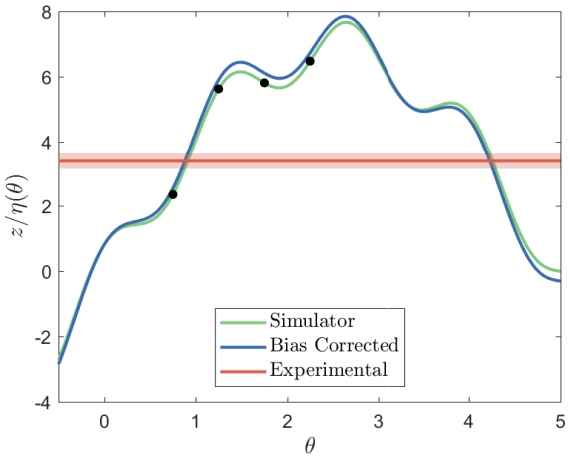

FIGURE 1. Simulator, model discrepancy and observational data (where the shaded region is $\pm\sqrt{V_o+V_m}$) for BHM numerical example. Where the initial simulator runs are $(\cdot)$.

$\{0.75, 1.25, 1.75, 2.25\}$. The observation $z = 3.39$ can be formed from two parameter 0.90 and 4.23 indicated by the cross-over in fig. 1.

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 a constant mean and Squared Exponential (SE) covariance functions (a fixed nugget term $\nu = 1 \times 10^{-8}$ [13]). The first, second and fourth waves are shown in fig. 2.

In the first wave (top panel of fig. 2) 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 parameter, around 1 and 2, being classed as non-implausible. By the second wave (top middle panel of fig. 2) 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 ($k = 4$) the code uncertainty has reduced across the space, and is lower than the remaining uncertainties in the non-implausible region; this took 16 simulator evaluations. The non-implausible set θ_{nI} at this wave clearly contain two regions around the solution 0.90 and 4.23.

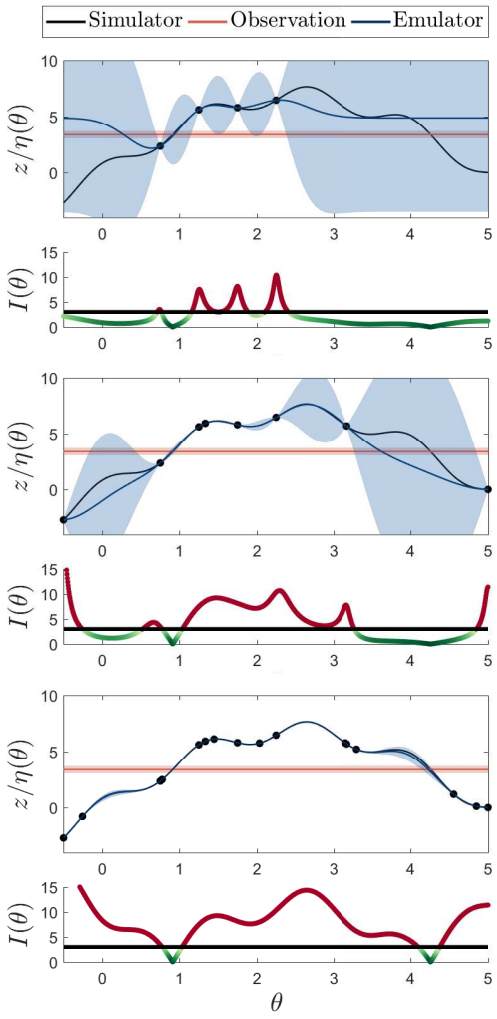

FIGURE 2. BHM waves $k = 1, 2, 4$ for the numerical example. The top part of each panel shows the observational data with $\pm\sqrt{V_o+V_m}$ shaded region against the simulator and emulator predictions (where the shaded regions indicates $\pm 3\sigma$), trained using the simulator runs $\eta(\theta^k)$ $(\cdot)$. The bottom part of the panel shows the implausibility $I(\theta_s^k)$ against the threshold $T = 3$, where green regions are non-implausible and red implausible. The top, middle and bottom panels show waves $k = 1, 2, 4$ respectively.

SEQUENTIAL DESIGN APPROACHES

Central to implementing BHM is generating and evaluating computer Design of Experiments (DoE)s. These provide the information required to construct emulators with which to assess and classify the parameter domain in a computationally efficient manner. As a result alternative DoE formulations can be used, as opposed to space-filled designs, such as GMLHD, which have been used in the original formulation of BHM. There are sev-

eral types of DoE that exist within the literature e.g. Monte Carlo sampling techniques, LHDs, Sobol sampling, Halton sampling and entropy (or information-based) sampling. In keeping with the brevity of this paper the reader is referred to [14] for an overview of these techniques. Two heuristic sequential based methods are explored in this paper with a view to move towards information-based DoEs. Two metrics, probability of non-implausibility and expected (un)improvement, adapted from the field of Bayesian optimisation and reformulated for BHM, provide criteria for selecting new simulator evaluations in a sequential manner and are explored in the following sections.

Probability of Non-implausibility

Probability of non-implausibility assesses the chance of a parameter combination being non-implausible given the observation and system uncertainties [15]. Mathematically this is the probability that $\theta \in \theta_{nI}$ if the mean prediction from the emulator lies within the uncertainty bounds, $D_{-,j}(x) \leq \mathbb{E}_j(\mathcal{GP}(x,\theta)) \leq D_{+,j}(x)$ as defined for the ith parameter combination eq. (10).

$$p(\theta_i \in \theta_{nI}) = \Phi\left(\frac{D_{+,j}(x) - \mathbb{E}\left(\mathcal{GP}_j(x,\theta_i)\right)}{V_{c,j}(x,\theta_i)^{-0.5}}\right) - \Phi\left(\frac{D_{-,j}(x) - \mathbb{E}\left(\mathcal{GP}_j(x,\theta_i)\right)}{V_{c,j}(x,\theta_i)^{-0.5}}\right) \quad (10)$$

Where $D_{+,j}(x)$ and $D_{-,j}(x)$ are the upper and lower non-implausible output bounds $z_j(x) \pm v_s\sqrt{V_{o,j}+V_{m,j}}$ with v_s defining the bound width, and $\Phi(\cdot)$ a standard Gaussian CDF. This variance scalar effectively behaves as the threshold in the implausibility metric and here is set as 3 due to Pukelsheim's 3σ rule.

The probability of non-implausibility is similar to the probability of improvement used in Bayesian optimisation. This heuristic when implemented in Bayesian optimisation is used to determine the probability of improving on the current minimum across a space [16]. In contrast the formulation in eq. (10) seeks parameter combinations that are likely to be within the output bounds $[D_{+,j}(x)\ D_{-,j}(x)]$, leading to the confident exclusion of parameter regions when the probability of being non-implausible is close to zero and the reverse when probability is close to one. The non-implausibility criteria is therefore defined as parameter combinations where $p(\theta_i \in \theta_{nI}) = 1$.

In sequential BHM each wave seeks to find the parameter combination with the largest probability less than one and to use this set as the next simulator evaluation. This reflects the belief that probability one states — with certainty given the bounds — that the parameter set output matches the output bounds, where the largest probability less than one (and greater than zero) will indicate a potential match which could be made certain either way by improving the code uncertainty of emulator prediction for that set. A stopping criteria can be formed similar to algorithm 1 where the process stops when the code uncertainty of the parameters with probability greater than zero is less than the observational and model discrepancy uncertainties.

Figure 3 demonstrates a selection of waves when probability of non-implausibility is implemented as part of a sequential BHM approach for the numerical example in fig. 1; with the same emulator mean and covariance functions and uncertainties. Between waves 1 and 5 (the top two panels of fig. 3) it can be seen that the algorithm spends simulator evaluations exploiting the nearby non-implausible region, with the next simulator evaluation for wave 6 being away from this area. The algorithm becomes more exploratory between waves 5 and 10, where the second non-implausible region is starting to be identified. Finally by wave 18 the stopping criteria has been met and the two non-implausible regions have been found. The approach requires more simulator evaluations, 18, than algorithm 1, 16. This is due to the probability of non-implausibility being a highly exploitative criteria, as shown by the numerous evaluations about the non-implausible regions.

By deriving the probability of non-implausibility, BHM can be defined as a subcategory of Approximate Bayesian Computation (ABC) [15]. Essentially this formulation becomes ABC with a uniform prior $p(\theta) \propto \mathbb{1}_{\theta\in\Theta}$ over the assessed parameter domain Θ and an acceptance kernel $\mathbb{1}_{\eta(\theta)\in[D_{+,j}(x)\,D_{-,j}(x)]}$ meaning the approximate posterior becomes,

$$p(\theta\,|\,z) \propto \begin{cases} 1 & \text{if } \theta \in \theta_{nI}, \\ 0 & \text{otherwise} \end{cases},$$

where a posterior probability of zero means an implausible parameter combination. This comparison allows BHM to gain useful properties from ABC, such as that ABC performs exact inference under uniform additive model discrepancy [17].

Expected (un)Improvement

Another heuristic with an improved balance between exploratory and exploitative objectives is expected (un)improvement. This proposed sequential design criteria is a development and reformulation of expected improvement utilised in Bayesian optimisation [18] combining the probability of matching observations within the uncertainty bounds with the expected magnitude of the improvement at a particular parameter combination.

To construct the criteria, (un)improvement must be defined; where improvement is typically $I(\theta) = \max(f_{min} - \eta(\theta), 0)$ in Bayesian optimisation [18]. This definition states that an improvement occurs when the simulator prediction is less than the

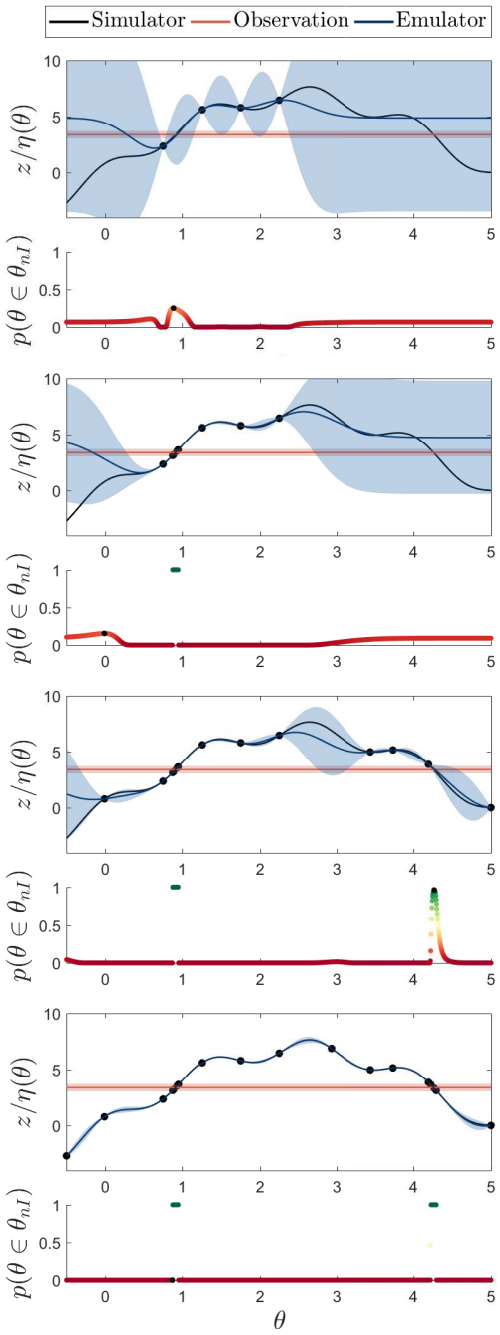

FIGURE 3. Sequential BHM using probability of non-implausibility for waves $k = 1, 5, 10, 18$ (from top to bottom panels) for the numerical example. The top part of each panel shows the observational data with $\pm\sqrt{V_o + V_m}$ shaded region against the simulator and emulator predictions (where the shaded regions indicates $\pm 3\sigma$), trained using the simulator runs $\eta(\theta^k)$ $(\cdot)$. The bottom part of the panel shows the probability of non-implausibility $p(\theta \in \theta_{nI})$, where $(\cdot)$ indicates the new simulator evaluation for the $(k+1)$th wave.

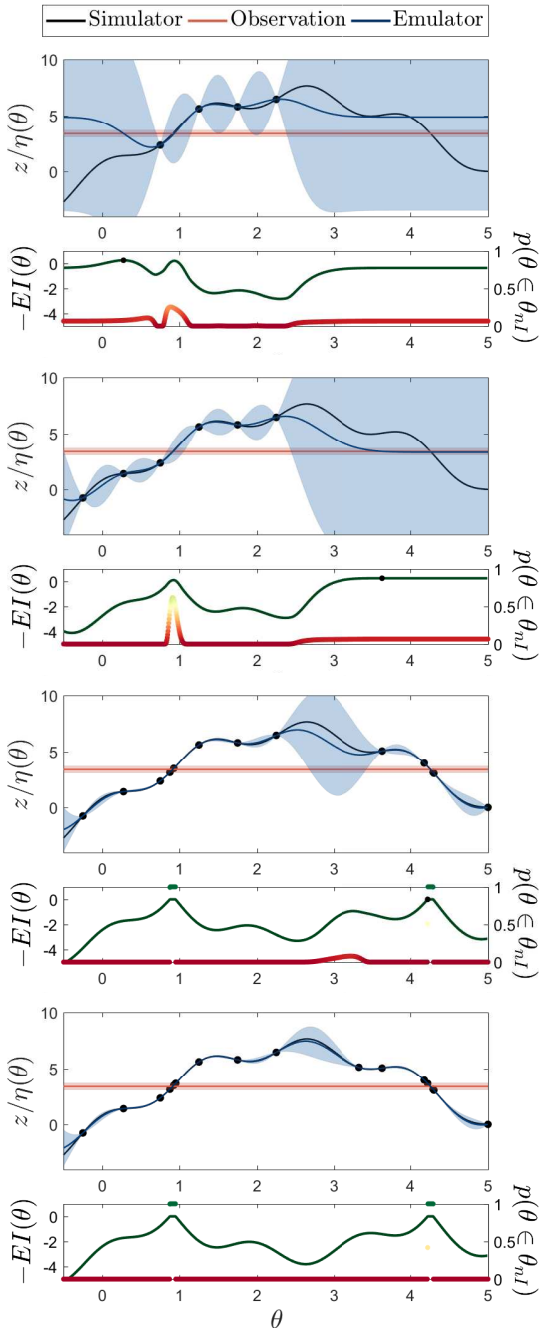

FIGURE 4. Sequential BHM using expected (un)improvement for waves $k = 1, 3, 10, 14$ (from top to bottom panels) for the numerical example. The top part of each panel shows the observational data with $\pm\sqrt{V_o + V_m}$ shaded region against the simulator and emulator predictions (where the shaded regions indicates $\pm 3\sigma$), trained using the simulator runs $\eta(\theta^k)$ $(\cdot)$. The bottom part of the panel shows the negative expected (un)improvement $-EI(\theta)$ and probability of non-implausibility $p(\theta \in \theta_{nI})$, where $(\cdot)$ indicates the new simulator evaluation for the $(k+1)$th wave.

current function minimum, with the improvement being zero when the simulator prediction is lower. In a BHM context the function minimum f_{min} is replaced by the observation with its defined uncertainty bounds. In this context the notion of improvement is not what is required, instead the search is for the 'smallest improvement' from the known bounded observations. In addition there are two improvement criteria as the observation is upper and lower bounded. This leads to the formulation of a criteria that will be zero or positive when a parameter is within the observation bounds and negative for the reverse. This sequential criteria is designed from taking the expectation of two (un)improvement criteria, where an (un)improvement occurs when the expected emulator prediction is below the lower bound $I_{lb}(\theta) = \max(D_- - \mathbb{E}(\mathcal{GP}(\theta)), 0)$ or greater than the upper bound $I_{ub}(\theta) = \max(\mathbb{E}(\mathcal{GP}(\theta)) - D_+, 0)$. The expected (un)improvement for all possible emulator values at a parameter combination is found by taking the expectation, which can be calculated in closed form for the lower and upper bounds in eqs. (11) and (12) respectively.

$$\mathbb{E}_{\eta\sim\mathcal{GP}(\theta)}(I_{lb}(\theta)) = \sqrt{V_c(\theta)}(\gamma_{lb}\Phi(\gamma_{lb}) + \Phi(\gamma_{lb})) \quad (11)$$

$$\mathbb{E}_{\eta\sim\mathcal{GP}(\theta)}(I_{ub}(\theta)) = \sqrt{V_c(\theta)}(-\gamma_{ub}\Phi(-\gamma_{ub}) + \Phi(\gamma_{ub})) \quad (12)$$

Where $\gamma_{lb} = (D_- - \mathbb{E}(\mathcal{GP}(\theta)))/\sqrt{V_c(\theta)}$ and $\gamma_{ub} = (D_+ - \mathbb{E}(\mathcal{GP}(\theta)))/\sqrt{V_c(\theta)}$ are the standardised distances between the bounds and mean emulator prediction. The expected (un)improvement criteria is the negative sum of eqs. (11) and (12) as defined in eq. (13) and takes the same units as the emulator output.

$$-EI(\theta) = -(\mathbb{E}_{\eta\sim\mathcal{GP}(\theta)}(I_{lb}(\theta)) + \mathbb{E}_{\eta\sim\mathcal{GP}(\theta)}(I_{ub}(\theta))) \quad (13)$$

The criteria can be combined with probability of non-implausibility to form a sequential BHM algorithm, where the approach follows that outlined previously (with the same non-implausibility and stopping criteria) with a different method for selecting new simulator evaluations. New runs are obtained for the parameter combination, with probability of non-implausibility less than one, where the expected (un)improvement ($-EI(\theta)$) is maximum.

Figure 4 presents a selection of waves from performing sequential BHM using expected (un)improvement for the numerical example with the same emulator mean and covariance functions and uncertainties. By wave 3 the method has begun exploring the parameter space with simulator evaluations concentrated at the observation bounds, as with probability of non-implausibility. Wave 10 demonstrates that the approach has explored the parameter space and begins to exploit locations that are likely to be plausible. At iteration 14 the algorithm has met the stopping criteria showing a greater efficiency than both probability of non-implausible and the space-filling approaches.

CONCLUSION

BHM is a methodology for performing parameter inference whilst accounting for model discrepancy. The technique involves an iterative procedure whereby an emulator is improved with new simulator evaluations. Current methods use a space-filling design to generate these new evaluation locations. In contrast, this paper has developed and explored the use of two metrics, probability of non-implausibility and expected (un)improvement as methods for performing a sequential DoE within BHM.

These metrics were assessed using a numerical case study in which it was shown that probability of non-implausibility performed worse than a standard space-filling design due to the metrics emphasis on exploitation with no regard to exploration. In comparison, expected (un)improvement was found to provide a better balance between exploitation and exploration, reducing the number of simulator evaluations from the standard space-filling approach.

Further research should conducted into information-based metrics, such as entropy search [19, 20]. These techniques seek to design criteria from information theory that maximises the expected information gain on the GP posterior, and should lead to a more efficient sequential approach. In addition, the outlined metrics should be applied to an experimental problem with a more complex simulator form, understanding the performance of this approach in higher dimensions. This will allow further comparison between the techniques.

REFERENCES

[1] Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A., 1997. "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments". In *Lecture Notes in Statistics*. pp. 37–93.

[2] Goldstein, M., and Paulo, R., 2012. "External Bayesian analysis for computer simulators". *Bayesian Statistics 9*(1996).

[3] Vernon, I., Goldstein, M., and Bower, R., 2014. "Galaxy formation: Bayesian history matching for the observable universe". *Statistical Science,* ***29***(1), feb, pp. 81–90.

[4] Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G., 2015. "Bayesian history matching of com-

plex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda". *PLoS Computational Biology,* ***11***(1).

[5] Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G., 2017. "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), aug, pp. 717–740.

[6] Edwards, N. R., Cameron, D., and Rougier, J., 2011. "Precalibrating an intermediate complexity climate model". *Climate Dynamics,* ***37***(7-8), pp. 1469–1482.

[7] Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K., 2013. "History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble". *Climate Dynamics,* ***41***(7-8), pp. 1703–1729.

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

[9] Fricker, T. E., Oakley, J. E., and Urban, N. M., 2013. "Multivariate Gaussian process emulators with nonseparable covariance structures". *Technometrics,* ***55***(1), pp. 47–56.

[10] Worden, K., Manson, G., and Fieller, N. R. J., 2000. "Damage detection using outlier analysis". *Journal of Sound and Vibration,* ***229***(3), pp. 647–667.

[11] Pukelsheim, F., 1994. "The three sigma rule". *American Statistician,* ***48***(2), pp. 88–91.

[12] Dette, H., and Pepelyshev, A., 2010. "Generalized latin hypercube design for computer experiments". *Technometrics,* ***52***(4), nov, pp. 421–429.

[13] Andrianakis, I., and Challenor, P. G., 2012. "The effect of the nugget on Gaussian process emulators of computer models". *Computational Statistics & Data Analysis,* ***56***(12), dec, pp. 4215–4228.

[14] Garud, S. S., Karimi, I. A., and Kraft, M., 2017. "Design of computer experiments: a review". *Computers and Chemical Engineering*.

[15] Holden, P. B., Edwards, N. R., Hensman, J., and Wilkinson, R. D., 2015. "ABC for climate: dealing with expensive simulators". *Handbook of ABC*, pp. 1–28.

[16] Snoek, J., Larochelle, H., and Adams, R. P., 2012. "Practical Bayesian optimization of machine learning algorithms". In Advances in neural information processing systems, pp. 2951–2959.

[17] Wilkinson, R. D., 2013. "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error.". *Statistical applications in genetics and molecular biology,* ***12***(2), pp. 129–41.

[18] Jones, D. R., Schonlau, M., and Welch, W. J., 1998. "Efficient global optimization of expensive black-box functions". *Journal of Global Optimization,* ***13***(4), pp. 455–492.

[19] Hennig, P., and Schuler, C. J., 2012. "Entropy search for information-efficient global pptimization". *Journal of Machine Learning Research,* ***13***, pp. 1809–1837.

[20] Chevalier, C., Ginsbourger, D., Bect, J., Vazquez, E., Picheny, V., and Richet, Y., 2014. "Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set". *Technometrics,* ***56***(4), pp. 455–465.