Efficiency and Reliability of Bayesian Calibration of Energy Supply System Models

Kathrin Menberg[1,2], Yeonsook Heo[2], Ruchi Choudhary[1]
[1] University of Cambridge, Department of Engineering, Cambridge, UK,
[2] University of Cambridge, Department of Architecture, Cambridge, UK,

Abstract

In this study, we examine the efficiency and reliability of a Bayesian calibration setup using temperature point measurements. Hamiltonian Monte Carlo sampling is found to be significantly more efficient with regard to convergence of the posterior distributions, which is assessed using different visual and quantitative measures. The examination of posterior realizations from different data sets and different prior distributions reveals that inference about model parameters is in general quite reliable, while learning about the magnitude of different error terms, such as model discrepancy and random errors, proves to be more difficult. Finally, predictive simulation results based on these inferred posterior distributions are generally in good agreement with measured data.

Introduction

Recent years have seen an increasing use of Bayesian calibration (BC) approaches in building energy modelling, as they allow consideration of different sources of uncertainty in both the model and input parameters (Heo et al., 2014; Li et al., 2016; Tian et al., 2016). Another advantage of BC approaches is that they allow inference about model parameters and predictive simulation, when a small amount of measured field data (approx. 10-30) is available for calibration (Omlin & Reichert, 1999). In the current literature, models are most often calibrated against aggregated outputs (such as monthly energy use) (Heo et al., 2012), where extreme model or parameter behaviour is often cancelled out due to data averaging. In this paper, we address this problem by examining cases, where point measurements, such as temperatures, are used as field observation points, instead of aggregated data as in most previous studies. In such cases, outliers might have a significant effect on the calibration results – especially when only a small number of data points are used for calibration.

In addition, Bayesian calibration or inference approaches typically employ Markov Chain Monte Carlo (MCMC) sampling strategies, which have a very low acceptance rate for the Metropolis-Hastings criterion and thus a slow convergence for multi-dimensional parameter spaces (Gelman et al., 2014). As building energy models commonly have many unknown parameters, the number of iteration steps required to achieve convergence for a high-dimensional parameter space is likely to affect the applicability of this calibration method in terms of computational time (Berger et al., 2016).

In this study, we address the problem of extensive computational time by for the first time employing Hamiltonian Monte Carlo (HMC) in the calibration framework and test its performance against standard random walk MCMC, which was employed by most previous studies. A comparison of different sampling methods requires assessment of the convergence speed of the samples towards the target distribution. Measures for evaluating the numerical convergence have so far not been applied to examine the reliability of calibration results in BC routines in the context of (building) energy models. Typically, the first several thousand of a vast number of performed simulation runs are rejected to allow for a certain level of convergence, but without systematic assessment. In this paper, we address the issue of convergence control by employing different visual and quantitative convergence measures for the Monte Carlo (MC) iterations. We also compare the applicability of the different convergence measures regarding the level of convergence required to obtain reliable calibration results.

We apply the Bayesian calibration framework developed by Kennedy and O'Hagan (2001) on a ground source heat pump system modelled in TRNSYS, and use measurements of the load side temperature to calibrate the model and to infer uncertain model parameters. Different field data sets are used in order to investigate the role of outliers and different field data trends in the BC approach. In addition, we examine the reliability of the BC framework for the first time in detail with regard to its ability to infer unknown model parameters and different model error terms, and predict model outcomes by posterior simulation.

Methodology

Bayesian calibration framework

In this automated calibration process, Bayesian inference is used to obtain the probability distributions for unknown model parameters while accounting for uncertainty in parameters, data and the model. The approach is based on Bayes' paradigm (eq.1), which relates the probability p of an event (or parameter value, θ) given evidence (or data, y), $p(\theta|y)$, to the probability of the event, $p(\theta)$, and the likelihood $p(y|\theta)$ (Gelman et al., 2014):

$$p(\theta|y) \propto p(\theta) \times p(y|\theta) \quad (1)$$

Eq. (1) allows us to combine our prior belief about an event and evidence about this event, i.e. measured data, to update our belief and quantify it in form of posterior probabilistic distributions.

Kennedy & O'Hagan (2001) formulated a comprehensive mathematical framework (KOH framework), which applies Bayes' paradigm to the model calibration process using the relationship in eq. 2:

$$y_f(x) = \zeta(x) + \varepsilon = \eta(x,\theta) + \delta(x) + \varepsilon + \varepsilon_n \qquad (2)$$

Here y_f are field observations; ζ is the true physical process that cannot be observed; ε represents the measurement error; $\eta(x, \theta)$ is the model outcome, which depends on the state variable x (e.g. outdoor temperature) and the unknown model parameter(s) θ; $\delta(x)$ is the discrepancy between the model and the true process, and ε_n is the numerical error term.

To obtain an approximation of the posterior probabilistic distribution of the unknown model parameters, repeated model evaluations with iterative sample draws are required. As most building simulations are computational expensive, it is more convenient to use an emulator instead of the original model. In accordance with previous studies (Kennedy & O'Hagan, 2001; Heo et al., 2012) we use Gaussian processes (GP) to emulate the simulation outcome $\eta(x, \theta)$ and the model discrepancy function $\delta(x)$. GP models are a generalization of nonlinear multivariate regression models and quantify the relation between individual parameters and the model outcome by a mean and covariance function. Typically, GP models are assigned a zero mean function and the covariance matrix for the emulator term $\eta(x, \theta)$ is specified in the form of (Higdon et al., 2004) (eq. 3):

$$\Sigma_{\eta(i,j)} = \frac{1}{\lambda_\eta}\exp\left[-\sum_{k=1}^{p}\beta_{\eta,k}\left(x_{ik}-x_{jk}\right)^2 - \sum_{k'=1}^{q}\beta_{\eta,p+k'}\left(\theta_{ik'}-\theta_{jk'}\right)^2\right] \qquad (3)$$

This formulation introduces several unknown hyper-parameters to the calibration process: the precision hyper-parameter λ_η and a set of correlation hyper-parameters β_η, with the number of β_η depending on the number of state variables x and unknown model parameters θ.

The covariance of the GP for the model discrepancy term is formulated in the same manner as for the emulator term, but simplified as it only depends on x. Thus, the hyperparameter λ_b represents the precision of the covariance of the model discrepancy term, and the hyper-parameters β_b, one for each variable x, specify the correlation strength and the smoothness of the model discrepancy function. The other two error terms, ε and ε_n, are included as unstructured error terms (see eq. (2)). The random error terms do not depend on x, and are correspondingly specified solely by precision parameters λ_e and λ_{en}.

These hyper-parameters are also uncertain terms in the calibration process. We assign to them prior distributions following suggestions made in previous studies (Higdon et al., 2004; Heo et al., 2012):

λ_η	Gamma (10, 10)
λ_b	Gamma (10, 0.3)
λ_e	Gamma (10, 0.03)
λ_{en}	Gamma (10, 0.001)
β_η	Beta (1, 0.5)
β_b	Beta (1, 0.4)

These are Gamma distributions, specified by a shape and scale parameter. One assumes that the error terms are small, and hence these priors are skewed towards large values. We define prior Beta distributions for the correlation hyper-parameters, β_η and β_b, expecting strong and smooth correlations across the contour state x. Indeed, priors on β_η would depend very much on the expected variations in the model outputs with respect to x. Priors on β_b, on the other hand, indicates the ability of the model discrepancy function to capture most of the variations in observations.

This calibration process updates the range of likely values for the calibration parameters and hyper-parameters, referred to as posterior distributions. For the interpretation of the posteriors one has to bear in mind that they refer to the standardised model outcome and do not reflect the absolute magnitude of the error terms. These posteriors are used to make posterior predictive simulations of the model outcome over new values of x in form of interpolation or extrapolation. An overview of the whole calibration process from parameter screening to predictive simulation outputs is depicted in Figure 1.

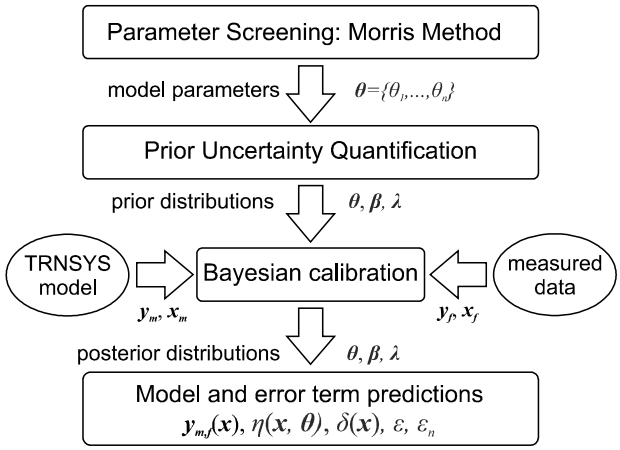

Figure 1: Overview of the calibration process.

Monte Carlo methods

Most studies applying Bayesian calibration use random walk Markov Chain Monte Carlo (MCMC) methods to sample iteratively from the joint posterior distribution, and employ the Metropolis-Hastings (MH) acceptance criterion to accept or reject the new samples, based on the posterior density of the current and the previous sample set (Gelman et al., 2014). This often poses a problem for multivariate, joint distributions, as the chances that one multidimensional set of random samples provides likely samples in all dimensions, i.e. for all parameters, are very low. Thus, the rejection rate in such a sampling process

may be quite high and accordingly it may take quite long to achieve convergence.

An alternative to MCMC, which offers a more efficient way to sample from a multivariate distribution, is Hamiltonian Monte Carlo. It is also called Hybrid Monte Carlo as it combines elements from stochastic Monte Carlo sampling with deterministic modelling of molecular particle dynamics (Duane et al., 1987; Neal, 2011). Instead of making one multivariate sample draw per iteration, HMC moves through the discretized multidimensional parameter space in several so-called leapfrog steps per iteration. Each sampled parameter is assigned an auxiliary momentum variable, which is updated at the same time, to enable a simultaneous evaluation of the parameter's position (i.e. value) and momentum. The momentum variable is related to the gradient of the posterior density. This process reflects the decomposition of a particle's total energy into potential and kinetic energy, and allows the identification of areas in the parameter space, where the amount of potential energy, and hence also the posterior density, are favourable. At the end of each iteration step, the MH acceptance criterion is applied to the vectors of sampled parameter values and momentum variables (Gelman et al, 2014).

We implement the Bayesian calibration framework with MCMC in Matlab, and use the STAN software (mc-stan.org) for the application with HMC. STAN employs a locally adaptive HMC with a no-U-turn sampler, which further enhances the performance of HMC by using the first iterations to optimize the tuning parameters, such as the discretization step size and the number of leapfrog steps per iteration (Carpenter et al., 2016).

Convergence measures

There exist different methods and measures to assess whether samples obtained from iterative simulations have converged to a representative approximation of the target (posterior) distribution. Typically, posterior simulation requires the evaluation of several sample sequences, called chains, with different random starting points in the parameter space. We employ an often-used visual method by inspecting the traceplots of the samples with increasing numbers of iterations and assessing the mixing between the different chains as well as stationarity of each individual chain, which indicate whether the chains cover the same common (target) distribution. In addition, we examine the autocorrelation plots based on the evaluation of the Markov chain criterion, which states that in a Markov chain the value of each new sample depends only on the previous sample. Thus, correlation for the samples within one chain should be low, but iterative sampling draws typically exhibit strong within-sequence correlations. While strong autocorrelation of posterior samples does not necessarily indicate a lack of convergence, it can have strong impact on the precision of predictive simulations (Gelman et al, 2014).

A frequently used quantitative measure to assess the convergence of iterative samples is the potential scale reduction factor, also known as $\hat{R}$ (Gelman & Rubin, 1992). We use this measure to assess the mixing of several chains by comparing the variance of samples within each chain, to the variance between the chains using the average of all samples within each chain. It is based on the consideration that for converged samples the covered distribution within each chain should be equal to the overall distribution covered by all chains and consequently all samples. Thus, a $\hat{R}$ value of 1.0 indicates very good convergence, and values below the threshold of 1.1 are typically assumed to indicate sufficient convergence for inference purposes. However, it should also be noted that proving convergence of iterative samples is not straightforward, and a $\hat{R}$ value of 1.0 does not guarantee that parameters are converging towards the target distribution. (see Gelman et al., 2014).

Data and modelling approach

For this study, we set up a model of a heat pump as part of a ground source heat pump system in TRNSYS and focus on its operation in cooling mode. We perform a sensitivity analysis on the model using Morris method (Morris, 1991) to identify the most influential model parameters for the load side outlet temperature as quantity of interest, as it was shown to provide reliable results with a low computational effort for this type of model (Menberg et al., 2016). Within the most important parameters, we select those that are uncertain (θ): load side fluid specific heat [kJ/kgK] (θ_1), rated cooling capacity of the heat pump [kJ/hr or W] (θ_2), source side flow rate in the heat pump [kg/hr] (θ_3), and load side flow rate in the heat pump [kg/hr] (θ_4).

All four uncertain parameters are assigned a normal prior distribution, which is a common choice for an informative or weakly informative prior (depending on the chosen variance). In an initial calibration run, all four parameters are assigned the same prior distribution $N(0.5, 0.16)$ with respect to their normalized ranges in (0,1). The parameter values for the specific heat of the cooling fluid and the rated cooling capacity can be estimated based on the documentation of the system specifications and are assigned a range of ±10%. As we are less certain about the true values of the flow rates, we assign a range of ±20% around the estimated mean values.

The load side inlet temperature of the heat pump represents an important input to the TRNSYS model and is selected as contour state variable x for the calibration process. Measured data for the inlet and outlet temperatures of the heat pump of the GSHP system of the Architecture Studio at the University of Cambridge is available as 15 min interval data for a period of two years. We select different data sets from the vast number of measurements available to be able to examine the influence of outliers and different data trends on calibration results. The data points were selected based on detailed statistical inspection of the total available data and refer to periods when the system was steadily operating at full or partial load capacity, but exclude any effects from turning the system on or off as measurements from the first two hours of each operation period are disregarded. By selecting these subsets of hourly data instead of using time-series data, the approach ignores time correlation of the measured data. The effects of time-dependency are expected to be very

small as system components have a very low thermal mass and their thermal inertia effect is usually not accounted for in the physical model.

Each calibration run constitutes a sequence of computer simulation results and field data, which consists of 10 measurement of the load outlet temperature y_f at corresponding inlet temperatures x_f (Figure 2). For the computer simulation outputs y_c, the heat pump model is evaluated at the same conditions x_c (= x_f), where field data is available. At each point x_c, the model is evaluated several times with varying values for the unknown model parameters θ using 40 Latin Hypercube samples for each θ covering a predefined parameter range.

The assembled field data points and the resulting computer outputs are displayed in Figure 2, which shows the different x ranges and different trends for the data sets A, B and C. Data set A contains field data with an almost linear trend that is entirely covered by the range of computer outputs, while data set B covers the same range of x, but has some significant outliers, which at the end of the computer output ranges. They are representative for cases, in which the general trend in the measured data agrees well with the simulation outputs, but random effects of different magnitudes impact the quality of measurements. Data set C consists of data for higher x values with few outliers, but a significant difference between field data and computer simulation results.

This data sets relates to scenarios, in which the general trend of field and simulation data is quite different, indicating that the model might not sufficiently represent some of the underlying physical effects. In our case, this data refers to summer days with quite hot temperatures in the heat pump inlet and outlet fluids, which are on the extreme ends of the specified part load curves of the heat pump model.

Results and Discussion

Convergence with MCMC and HMC

The number of iterations needed for sampling from the posterior distribution is significantly lower with Hamiltonian Monte Carlo (HMC) than with random walk Markov Chain Monte Carlo (MCMC). On average, for the heat pump model in this study, 1000 HMC iterations are sufficient to achieve a very good convergence with $\hat{R}$ values around 1.05 for all calibration parameters and most of the tuning parameters related to the locally adaptive HMC process.

The time needed for each single iteration, however, is significantly higher with HMC. One MCMC run took about 0.17s, compared to one HMC iteration of ca. 4.8s. This is partly due to the more complex calculations within an HMC iteration, but also possibly related to manner in which STAN is set up. The algorithms used in STAN for calculating the Cholesky decomposition of the covariance matrices of the Gaussian processes are less efficient than in MATLAB, which was used for the MCMC evaluations. Nevertheless, the lower number of iterations required for convergence with HMC more than compensates this inefficiency, as MCMC often requires more than 50,000 iterations to achieve good convergence for our calibration runs. In addition to the different sampling strategies, the number of observations and uncertain parameters, as well as the type of data set and prior distributions, also significantly influence the number of iterations needed to achieve convergence. Another important issue is of course the efficiency of the Matlab or STAN code itself, which is not considered in this example.

To assess the convergence of the calibration runs with MCMC and HMC we employ traceplots for parameter values, autocorrelation plots and the $\hat{R}$ value. Figure 3 shows the corresponding graphs for the GP precision hyperparameter of the random error λ_e, which often convergence very slowly, from an evaluation of data set B with 10,000 MCMC iterations. The first 5000 iterations are discarded as burn-in and are not shown in the traceplot and autocorrelation graph.

The traceplots for six sample chains show a good mixing between the chains and a stationary trend of parameter values within in the chains (Figure 3). Thus, these plots suggest that the chains cover the same distribution for random error λ_e in the GSHP model, and accordingly one can assume good convergence.

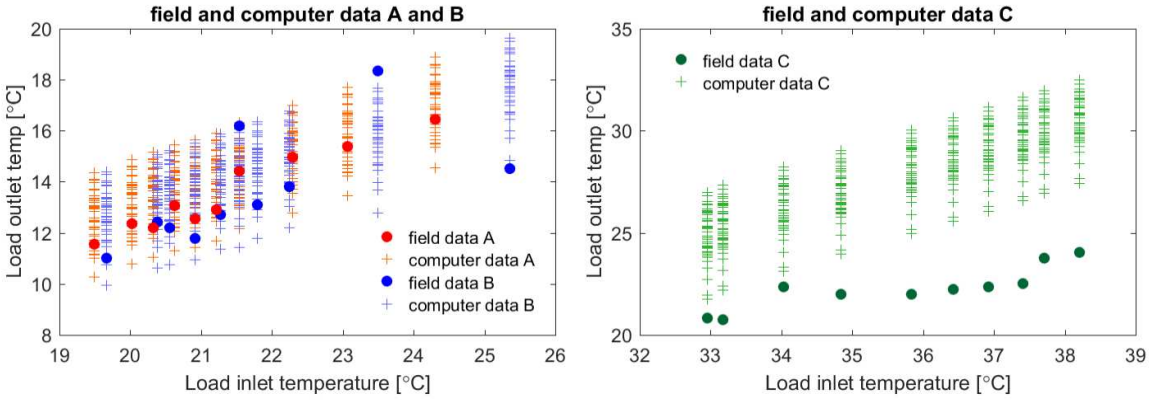

Figure 2: Measured field data and simulated computer data for three data sets A, B, and C showing the different temperature ranges and trends of each data set.

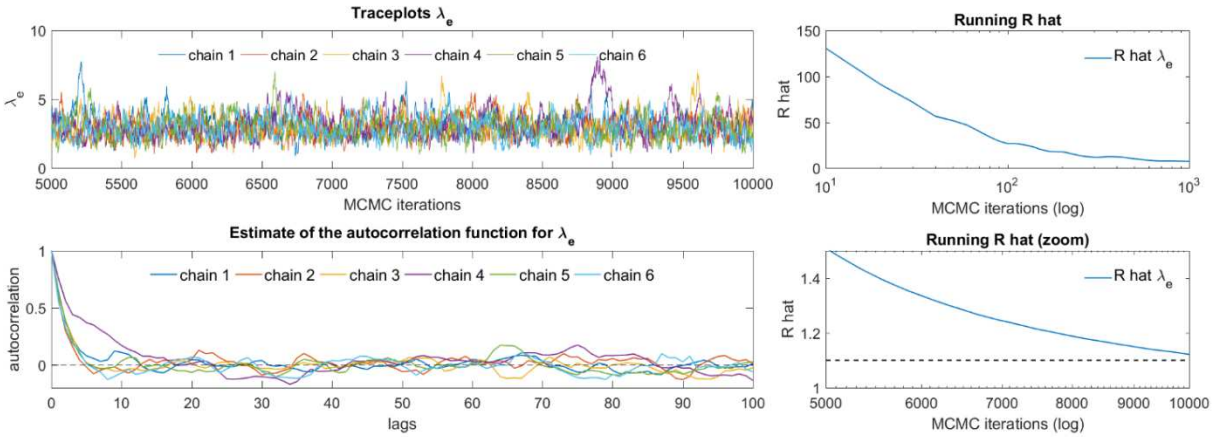

Figure 3: Example for different convergence measures applied to the posterior results for the random error parameter λ_e with MCMC sampling.

The autocorrelation plot shows a distinct decrease for the correlation coefficient with increasing lag size and the bottom plot of Figure 3 shows that coefficients for a lag size > 15 vary mostly between 0.2 and -0.2, which means that the samples are mostly independent. The evolution of the $\hat{R}$ value with increasing iteration number however indicates a much slower trend towards convergence (note the logarithmic x scale in the right-hand graphs in Figure 3). Indeed, the commonly used threshold of $\hat{R} = 1.1$ is not yet reached after 10,000 MCMC iterations. Here, the use of a numerical threshold has a clear advantage over the visual inspection of sample chains for defining that a certain level of convergence is achieved or not yet reached.

On the other side, this also raises the question regarding the level of convergence ($\hat{R} < 1.1$ or visually mixed chains) that is required in order for the calibration to provide reliable results. We also question the extent to which lack of 'perfect' convergence with $\hat{R} < 1.1$ influences calibration results. Indeed, in this study even calibration runs with a large number of MCMC iterations (>50,000) rarely achieve a $\hat{R}$ value of 1.1, especially with regard to parameters related to the error terms, which results in a huge computational effort to reach the recommended threshold.

Therefore, we compare the calibration results, with regard to posterior distributions of uncertain parameters and posterior predictive simulations, of the not fully converged MCMC evaluation from Figure 3 against converged HMC results with $\hat{R} \approx 1.0$. The comparisons are shown in Figure 4. The mode values of posterior distributions of uncertain parameters θ_1, θ_2 and θ_4 for both converged and unconverged results shift into the same direction, which agrees with our expectations. The posterior distributions of θ_3 show no change from the prior distribution, which indicate either that the prior mode was already close to the true value, or that no inference can be made about this parameter.

The interquartile range of the converged θ posterior distributions (θ_1: 0.17, θ_2: 808, θ_3: 149, θ_4: 50) is in general smaller than for the unconverged results (θ_1: 0.22, θ_2: 1060, θ_3: 215, θ_4: 60). This indicates that convergence helps reduce the remaining uncertainty associated with the calibration parameters. This difference in uncertainty is also reflected in the plots showing the posterior predictive simulations for the emulator $\eta(x, \theta)$ and the model outcome $y(x)$, which show smaller standard deviations for the fully converged results. The posterior distributions for the hyperparameters λ and β are almost identical in both cases (not shown). Similar patterns can be observed for comparisons between the corresponding converged and unconverged results from data set A and C (not shown). The overall similarity between the posterior distributions and prediction results from converged and not fully converged calibration processes suggests that a value of $\hat{R} \approx 1.0$ is not strictly necessary to obtain valid calibration results. Finally, the decision about a level of convergence should be made based on the desired accuracy of the results for a specific calibration exercise.

In general, convergence issues in this framework seem to occur most often for hyperparameters relating to error terms. This observation may be linked to the fact that priors on random errors are generally set to be quite low – not because random errors are always small, but because modellers often do not have good intuition about these.

Although HMC is clearly more efficient than MCMC, using HMC does not guarantee good convergence. In fact, some calibration runs with HMC exhibit $\hat{R}$ values that stay constant at a high level or even increase with further iterations. Such increasing $\hat{R}$ values indicate a diverging trend in the individual sampling chains, when for instance one chain becomes trapped in a certain area of the parameter space, while the remaining chains continue to explore other areas of the parameter space. These trapped chains are reflected by long, rather flat traceplot lines, and may relate to the fact that the HMC sampler is more susceptible to failures in multi-modal distributions than MCMC (Neal, 2011). This behavior was mostly observed for calibration exercises where one (or more) of the posteriors for the error terms were more than an order of magnitude off the defined prior distribution. Repeated calibration runs with identical settings though might ultimately lead to valid results. In addition, one should also remember that a $\hat{R}$ value close to one does not necessarily mean that the posteriors converge toward the target distribution.

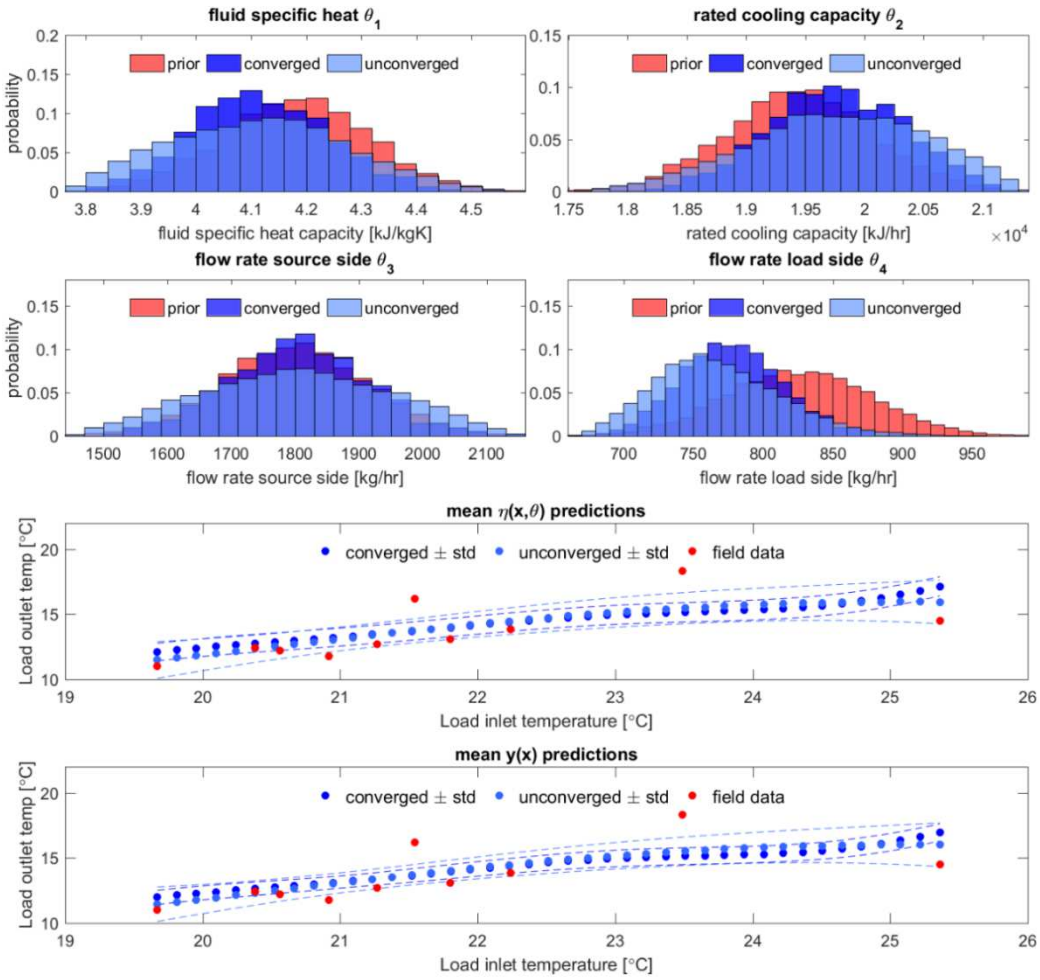

Figure 4: Posterior distributions for calibration parameters and hyperparameters, and predictions from inferred model parameters from converged HMC and unconverged MCMC analysis with data set B.

Inference about calibration parameters

One of the main objectives of the KOH calibration framework is to make inference about the uncertain model parameters. Ideally, the inferred posterior distribution should also result in the reduction of uncertainty in parameter values. As the setup is Bayesian, prior distributions play a significant role in the outcomes (posterior distributions). The left-hand column in Figure 5 shows the posterior distributions for the four calibration parameters assuming normal prior distributions $N(0.5, 0.16)$, which represent weakly informative priors with a rather large spread.

In order to investigate the influence of priors and test the identifiability of the model calibration parameters, we repeat the evaluation of the calibration process with a slightly different set of priors for all θ. For θ_1 we now assign a prior distribution with the same mean value of 0.5, but a smaller variance. The new prior for θ_2 has a higher mean value with a smaller variance than before. θ_3 is assigned a higher mean value and a higher variance. For θ_4, the variance of the prior distribution is increased, while the mean value is kept at 0.5.

For comparison, results are displayed in the right-hand column of Figure 5. The general trend of posterior distributions remains unchanged – especially in terms of the location of the posterior mode value. Small variations in the mode values are due to the fact that the posteriors are a compromise between the prior and the data.

Assigning smaller prior variances leads to more distinct posterior mode values (see Fig. 4 right-hand side θ_1 and θ_2) and consequently enables more exact inference about the unknown model parameters. On the other hand, prior distributions with a larger parameter range allow the calibration process to explore a larger range and results in slightly different posterior mode values due to the difference in prior variance as observed in θ_4 (bottom Fig. 4). θ_4 is also the only parameter that shows a significant reduction in posterior variance of 0.05 and 0.10, respectively, meaning that through the posterior inference we become more certain about the mode value of the load side flow rate. The source flow rate, θ_3, seems to be entirely unidentifiable with the current calibration setup, as both posteriors exactly mimic the assigned prior distributions.

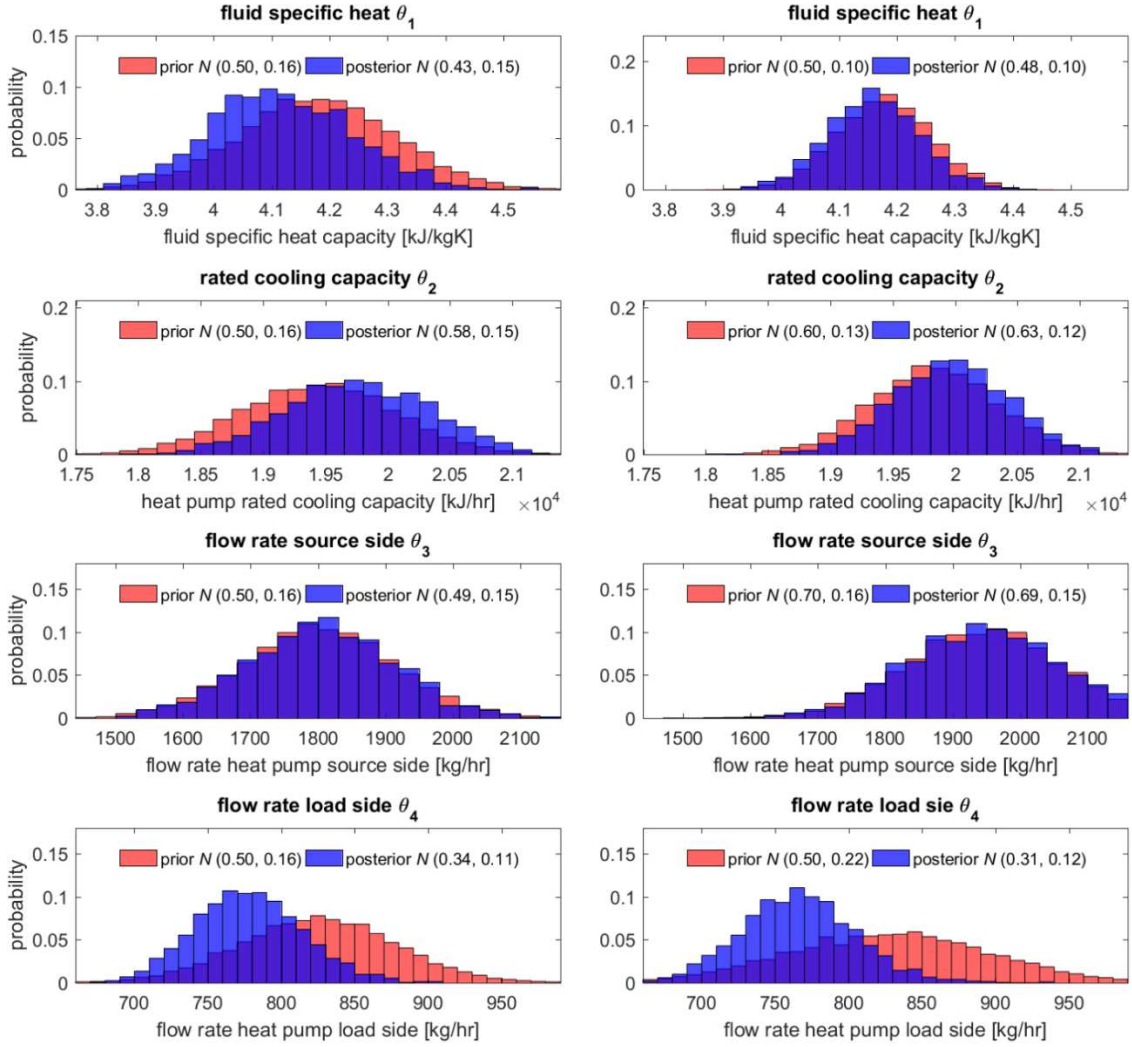

Figure 5: Posterior distribution for calibration parameters given different normal prior distributions for data set B. The characteristics given for the posterior distributions assume that they also follow a normal distribution and refer to the normalized parameter ranges.

In general, inference about mode values of the model parameters provides physically reasonable results. The posterior mode value for the specific heat of the cooling fluid, θ_1, is close to the value for water, which is in line with our expectations. The same applies to the posterior mode of rated cooling capacity, θ_2, which confirms the information from the technical data sheet.

Regarding the flow rates, θ_3 and θ_4, we do not have enough information about the true value to draw any conclusions about the correctness of the posteriors. The unidentifiability of θ_3, however, suggests that the uncertainty in this parameter may be subsumed by other (hyper) parameters. As the result of the Bayesian calibration process is a multivariate joint distribution for all unknown model and hyper-parameters, the posteriors of the θs are correlated and they have to be carefully interpreted together with the posteriors for the discrepancy and error terms.

Inference about model discrepancy term

In addition to inference about unknown model parameters, the KOH calibration also offers the opportunity to assess the magnitude of model discrepancy. Model discrepancy can be most simply described as the inability of the model to represent the physical process faithfully. The left-hand column in Figure 6 shows the prior and posterior distributions of the model discrepancy precision hyperparameter λ_b for the data sets A, B, and C. As these posteriors were inferred based on the standardized model outcome, the precision hyper-parameter does not directly reflect the absolute magnitude of the model discrepancy, but depends on the variance in the original model outputs.

For data sets A and B, the priors and posteriors are identical, which could indicate that no inference about the model discrepancy is possible. Therefore, we apply a different prior, GAM (10, 1), with lower values to λ_b, but the lack of change in the posterior indicates that inference about λ_b with these data sets is not possible (Figure 6, right-hand column). The right-hand graph for data set B

shows a very small shift in the posterior mode, which is may be spurious as there is no consistent trend of shifts in comparison to with the case with the first prior and additional evaluations with a broader range of priors (results not shown). This un-identifiability of the model discrepancy is likely to affect the reliability of the inference about the unknown model parameters as they are inherently linked through the joint multivariate distribution. However, both posterior distributions for λ_b from data set C show a significant shift to lower λ_b values than expressed by the prior belief. As the covariance of the Gaussian process is defined by the inverse of λ_b (eq. 3), this shift indicates a larger model discrepancy function than expected.
This observation is in accordance with our expectation about the model discrepancies based on the deviation of the model outcomes from the measured data for data set C in Figure 2. As this deviation takes the form of a rather constant offset in one direction, the calibration process assigns this to the model discrepancy function. On the contrary, deviations of measured and modelled system behaviour in varying directions shown in dataset B are absorbed by the random error term, as we chose priors on the beta hyper-parameters that reflect very rather smooth functions over x for the emulator and the model discrepancy.
It is also worth mentioning that the other two error terms, the measurement error λ_e and numerical error λ_{en}, are much larger for data sets A and B than for C (not shown), although the measurement and numerical procedures are identical for all the data. This raises the question of a potential confounding between the individual error terms due to the lack of identifiability of different error sources. A detailed discussion of the difficulties associated with identifying the model discrepancy term correctly is provided by Brynjarsdóttir & O'Hagan (2014). In addition, in the current calibration framework, the model discrepancy is defined as a structured error term, while the measurement and numerical errors are treated as random errors. Consequently, any structured error will be accounted for in the model discrepancy term, independently of its original source.

Posterior predictive simulation

Another main objective of calibration processes is the predictive simulation of the model outcome at new contour states (x) based on the posterior distributions of the model parameters and Gaussian process (GP) hyperparameters.
Given these GP hyperparameters, we first compare the posterior predictive results of the emulator term for the different data sets (Figure 7, left-hand column). The emulator based on data set A indicates an almost linear increase in load outlet temperature with increasing load inlet temperature following the trend given by the simulation outputs. The variation around the mean emulator output is caused by the variance of the posterior distribution of the calibration parameters (see Figure 5). It is important to note that the displayed standard deviation is the same across all x values, as the calibration parameters that affect the variation around the mean ($\theta_{1\text{-}4}$, λ_e, and λ_{en}) are independent of x (Figure 7).

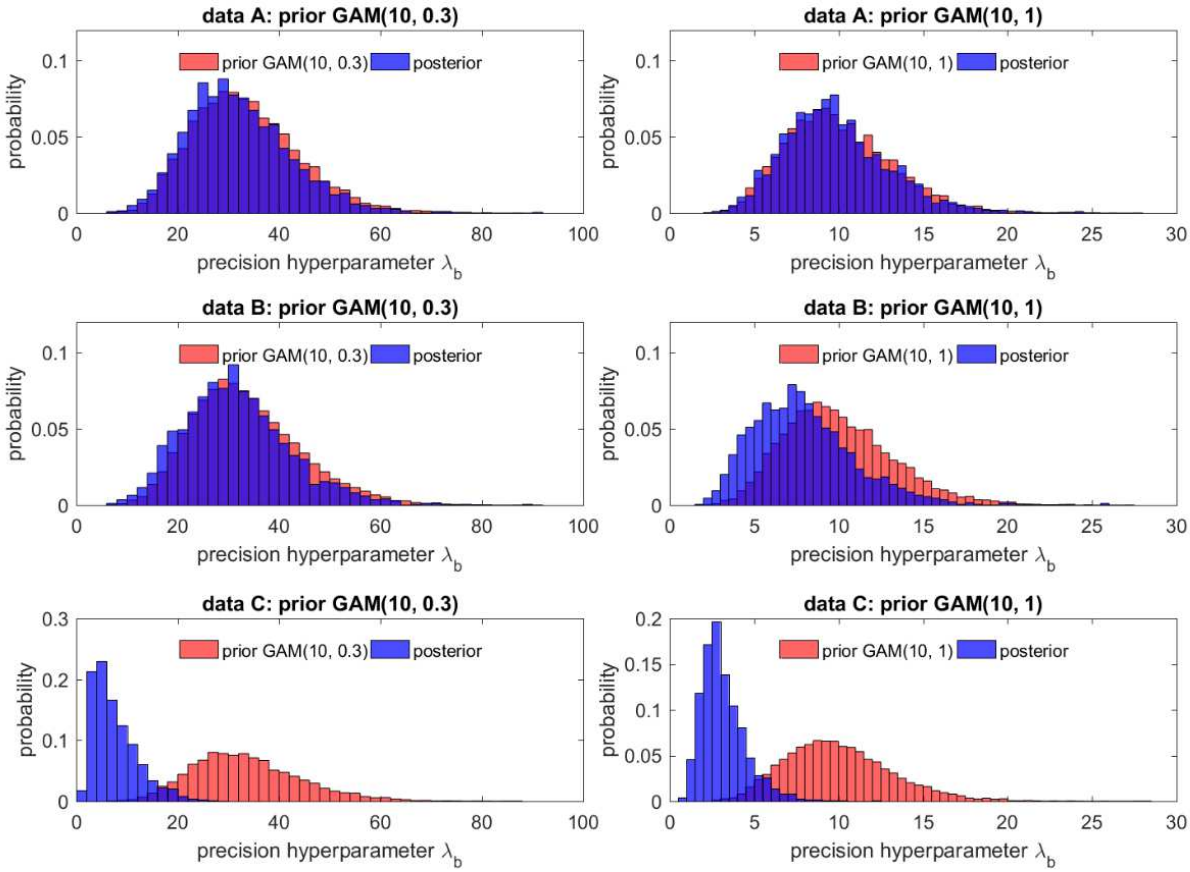

Figure 6: Posterior distribution for λ_b given different gamma prior distributions for data sets A, B, and C.

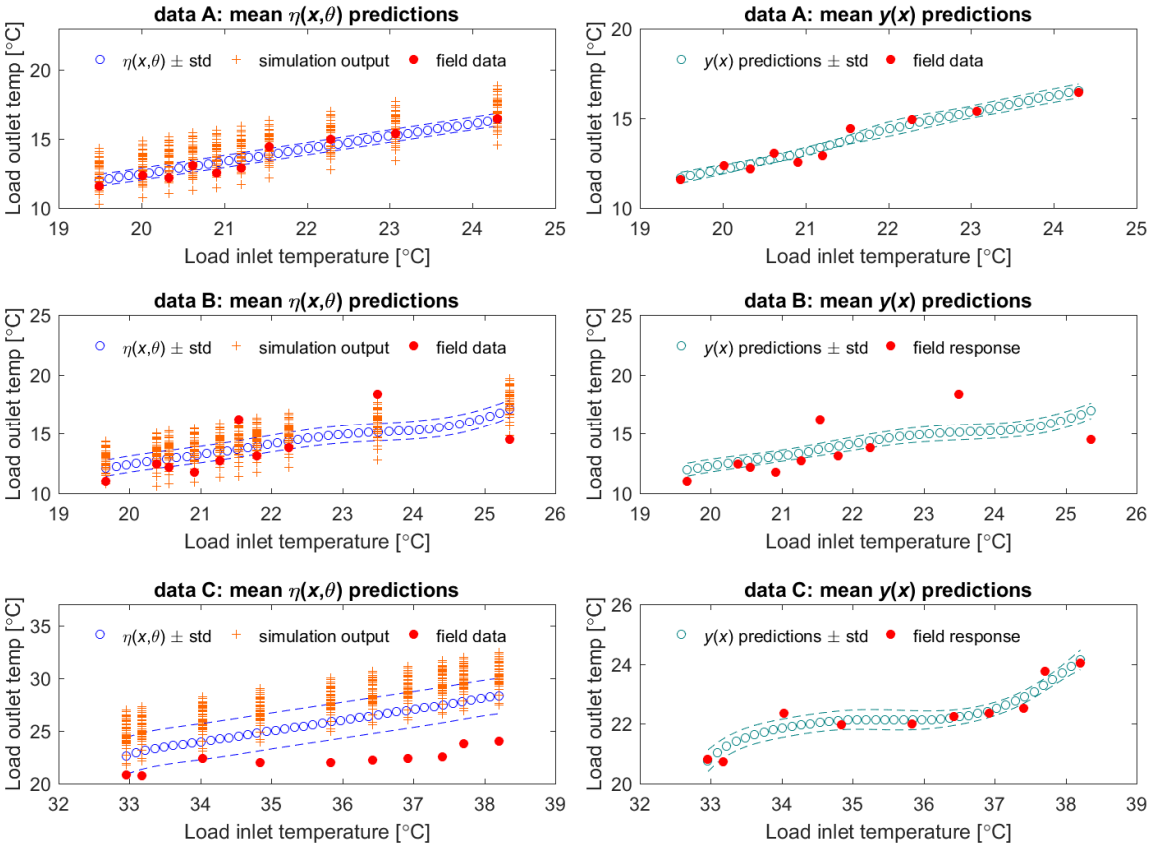

Figure 7: Posterior predictions for the emulator term and the model outcome for data sets A, B, and C.

The emulator of data set B produces a more curved shape, in particular in areas, where coverage of the x space is poor, such as for high x values. Given the significant gaps in the input data the calibration algorithm has less data points to rely upon, it tries to best match the few outlier points instead of following a linear trend as expected given our knowledge about the physical model.

Thus, larger gaps in combination with outliers play a significant role for the reliability of the emulator predictions, as the interpolation potential of the approach is limited. The trend of the emulator predictions for data set C is again almost linear and in line with our expectations, but shows a much wider range of uncertainty in comparison to the other cases.

Posterior predictions for the model outcome $y(x)$ shown in the right-hand column incorporate outcomes from predictions from the emulator and the inferred information about the model discrepancy (through λ_b and β_b). As mentioned above, inference about the model discrepancy for data A and B is very limited, which results in a close similarity between emulator and $y(x)$ predictions. Yet, accounting for this very small model discrepancy still leads to a visible reduction in the standard deviation of $y(x)$, especially in areas with dense x data points.

The posterior predictions based on data set B also provide a narrow range for potential $y(x)$ values. Some of the field data points are far outside the standard deviation, which means that they are not captured by the model discrepancy term. This is because the rigid structure of the model discrepancy term defined by the correlation hyperparameter β_b leads to a rather smooth discrepancy function that cannot capture the frequent deviations from the mean in both directions. Instead, these outliers are accounted for in the random error term λ_e.

The predictions for $y(x)$ based on data set C show the importance of including the model discrepancy term in the calibration and prediction process. The model discrepancy function allows for a full compensation of the offset between the linear emulator output and the measured field data, which results in a very good fit of $y(x)$ to the data with a low standard deviation.

Conclusions

The comparison of MCMC and HMC sampling strategies within the calibration framework shows that HMC enables much faster convergence, and thus reduces the number of iteration runs required leading to a significant improvement of computational efficiency. From the comparison of different convergence measures, we conclude that visual convergence measures require more experience to assess the level of convergence, while quantitative measures, such as the $\hat{R}$, can be easily interpreted. However, the comparison of fully converged ($\hat{R} \approx 1.0$) and not fully converged ($\hat{R} < 1.5$) results showed that calibration runs with $\hat{R}$ values larger than the typically used threshold can still hold reasonable and reliable results, albeit with slightly larger posterior spreads. Furthermore, it is important to remember that even a $\hat{R}$ value of 1.0 is no guarantee for convergence towards the target distribution

In general, the best way to assess convergence is probably a combination of quantitative and visual measures, and a careful examination of the posterior distribution for all (hyper-)parameters with regard to their underlying physical meaning and plausibility.

The type of prior distribution for unknown model parameters should be carefully considered, as a tight prior might prevent the posterior from converging towards the true parameter value. On the other hand, wide and uninformative priors can lead to wider posterior distributions with rather indistinct mode values. Ultimately, Bayes' paradigm leverages prior knowledge with additional evidence provided by data to infer uncertain parameters. Thus, a careful evaluation of prior knowledge or belief is needed, and in case of lacking or absent prior information and sufficient data points, other databased methods, such as maximum likelihood estimation may provide better options.

Regarding the inference about the model discrepancy function two of the three data sets show rather indistinct results, which may be related to the dominating effect of other error terms. In addition, the un-identifiability of the error terms reduces the reliability of the inference about all other model and hyper-parameters as they are all inherently linked through the joint multivariate distribution. Thus, more detailed analysis is needed to quantify the effect of such unidentifiable parameters on calibration results. For the third data set, a significant model discrepancy is identified in all evaluations, which corresponds well with our expectations based on the field and simulation data. However, more work is needed regarding the potential influence and identifiability of the correlation hyperparameters β_b and the potential confounding between the different structured and non-structured error terms.

The results for posterior predictive simulations with the three different data sets show that good coverage of the contour state space is a crucial factor, with data gaps leading to large variances in the prediction outcomes. Accounting for structured and non-structured error terms in the calibration framework allows for compensation of model discrepancy and outliers. This then leads to a good fit of the predictions to the field data with a narrow band of uncertainty around the average predictions. However, calibration results do not necessarily represent the 'true' parameter value or magnitude of the different error terms. The results of the described calibration process are joint posterior distribution, which typically exhibit strong correlations and dependencies between the individual distributions, and may lead to confounding between parameters with similar physical meaning or different error terms.

Acknowledgements

This study is supported by EPSRC grant (EP/L024452/1): Bayesian Building Energy Management (B-bem).

References

Berger J, Orlande HR, Mendes N, and S. Guernouti (2016). Bayesian inference for estimating thermal properties of a historic building wall. *Building and Environment 106,* 327-339.

Brynjarsdóttir, J. and A. O'Hagan (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems 30,* 114007.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li and A. Riddell (2016). Stan: A probabilistic programming language. *Journal of Statistical Software,* (in press).

Duane, S., Kennedy, A. D., Pendleton B. J. and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B 195,* 216-222.

Gelman, A., Carlin, J.B., Stern, H.S, Dunson, D.B., Vehtari, A., and D.B. Rubin (2014). Bayesian Data Analysis. 3rd Ed. CRC Press, Boca Raton, US.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7,* 457-472.

Heo, Y., Choudhary, R. and G. Augenbroe (2012). Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings 47,* 550-560.

Heo Y, Augenbroe G, Graziano D, Muehleisen RT, and L. Guzowski (2014). Scalable methodology for large scale building energy improvement: Relevance of calibration in model-based retrofit analysis. *Building and Environment 87,* 342-350.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo J. A. and R. D. Ryne (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing 26(2),* 448-466.

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

Li, Q., Augenbroe G. and J. Brown (2016). Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. *Energy and Buildings 124,* 194-202.

Menberg, K., Yeonsook, H., and R. Choudhary (2016). Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. *Energy and Buildings 133,* 433-45.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics 33,* 161-174.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2,* 113-162.

Omlin, M., and P. Reichert (1999). A comparison of techniques for the estimation of model prediction uncertainty. *Ecological Modelling, 115,* 45-59.

Tian W, Yang S, Li Z, Wei S, Pan W, and Y. Liu (2016). Identifying informative energy data in Bayesian calibration of building energy models. *Energy and Buildings* 119, 363-376.