



Bayesian calibration of building energy models with large datasets



Adrian Chong^{a,*}, Khee Poh Lam^{a,b}, Matteo Pozzi^c, Junjing Yang^a

^a Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566, Singapore

^b Center for Building Performance and Diagnostics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

^c Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:

Received 29 May 2017

Received in revised form 3 August 2017

Accepted 24 August 2017

Available online 31 August 2017

Keywords:

Building simulation

Bayesian calibration

Uncertainty analysis

Hamiltonian Monte Carlo

No-U-Turn Sampler

ABSTRACT

Bayesian calibration as proposed by Kennedy and O'Hagan [22] has been increasingly applied to building energy models due to its ability to account for the discrepancy between observed values and model predictions. However, its application has been limited to calibration using monthly aggregated data because it is computationally inefficient when the dataset is large. This study focuses on improvements to the current implementation of Bayesian calibration to building energy simulation. This is achieved by: (1) using information theory to select a representative subset of the entire dataset for the calibration, and (2) using a more effective Markov chain Monte Carlo (MCMC) algorithm, the No-U-Turn Sampler (NUTS), which is an extension of Hamiltonian Monte Carlo (HMC) to explore the posterior distribution. The calibrated model was assessed by evaluating both accuracy and convergence.

Application of the proposed method is demonstrated using two cases studies: (1) a TRNSYS model of a water-cooled chiller in a mixed-use building in Singapore, and (2) an EnergyPlus model of the cooling system of an office building in Pennsylvania, U.S.A. In both case studies, convergence was achieved for all parameters of the posterior distribution, with Gelman–Rubin statistics \hat{R} within 1 ± 0.1 . The coefficient of variation of the root mean squared error (CVRMSE) and normalized mean biased error (NMBE) were also within the thresholds set by ASHRAE Guideline 14 [1].

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

BEM is increasingly being used for the analysis and prediction of building energy consumption, M&V and the evaluation of ECMs. To ensure the reliability and accuracy of an energy model, model calibration has been recognized as an integral component to the overall analysis [40]. Calibration can be viewed as the process of tuning model parameters until the simulation predictions match the observed values reasonably well. However, models are only as accurate as the inputs provided and detailed information is seldom available because it may be prohibitively expensive or even impossible to measure every tuning parameter of the model. Consequently, calibrating these models with limited data can often lead to over-parameterization and equifinality (i.e., the model parameters are not uniquely identifiable) [2]. Additionally, buildings are made up of complex systems interacting with one another and thus no single model is beyond dispute. Therefore, it is clear that there is a need for a calibration framework that is able to account for uncertainties in the modeling procedure [3]. Incorporating uncertainty would also allow risk to be better quantified. For instance, a risk-

conscious decision-maker would prefer an ECM that yields a higher probability of guaranteed savings while a risk-taking decision-maker would prefer an ECM that yields the highest expected value [18].

Most calibration approaches that have been proposed are manual approaches that require the energy modeler to iteratively adjust individual parameters until a calibrated solution is achieved [9]. Westphal and Lamberts [48] used sensitivity analysis to first identify influential parameters, which were then adjusted to match simulation output to measured data. Through a case study, the approach was shown to be able to achieve a 1% difference between simulated annual electricity consumption and actual consumption. Using information from walk-through audits and end-use energy measurements, Pedrini et al. [35] carried out monthly calibrations and significantly reduced the difference between simulated annual electricity consumption and actual consumption. Tools such as graphical plots [40,28] and version control [39] have also been shown to be useful aids for guiding the calibration process. The main advantage of manual approaches is that model parameters are adjusted based on heuristics that are based on the expertise of an experienced modeler. However, these approaches are time consuming and labor intensive. They also rely heavily on the skills

* Corresponding author.

E-mail address: bdgczma@nus.edu.sg (A. Chong).

Nomenclature

Abbreviations

BEM	building energy model/modeling
COP	chiller coefficient of performance
CVRMSE	coefficient of variation of the root mean squared error
ECM	energy conservation measure
GP	Gaussian process
HMC	Hamiltonian Monte Carlo
IoT	Internet of things
KL	Kullback–Leibler
M&V	measurement and verification
MCMC	Markov chain Monte Carlo
NMBE	normalized mean biased error
NUTS	No-U-Turn Sampler
RWM	random walk metropolis

Physical quantities

\dot{m}_{chw}	chilled water mass flow rate [kg/h]
\dot{Q}_{load}	cooling coil load [W]
\dot{V}_{chw}	chilled water flow rate [m ³ /s]
\dot{V}_{frac}	fraction of peak chilled water flow rate [-]
$T_{chw,in}$	chilled water inlet temperature [°C]
$T_{chw,set}$	chilled water setpoint temperature [°C]
$T_{cw,in}$	condenser water inlet temperature [°C]

Uppercase Roman letters

\hat{R}	Gelman–Rubin statistics
D	dataset
J	information or Kullback–Leibler divergence
L	number of leapfrog steps
N	$= n + m$
P	percentile
Q	sample quality
S	sampling schedule

Lowercase Roman letters

c	number of bins or categories after discretizing
m	number of simulation data
n	number of observed values
p	number of input factors
q	number of calibration parameters
r	number of attributes
s	starting sample size
t	calibration parameters
x	input factors
$y(x), y$	observed output
z	$= [y_1, \dots, y_n, \eta_1, \dots, \eta_m]$

Greek letters

$\delta(x)$	discrepancy between simulation predictions and observed output
$\epsilon(x)$	observation errors
$\eta(x, t), \eta$	simulator output
λ	variance hyperparameter of GP model
μ	mean value of elementary effects
μ^*	absolute mean value of elementary effects
ρ	$= \exp(-\beta/4)$
Σ	covariance matrix
σ	standard deviation
θ	uncertain parameters

Superscript

δ	discrepancy term
η	simulator
f	field data
S	simulation data
T	transpose of a matrix

Subscript

δ	discrepancy term
η	simulator
i, j, k	parameter or variable index
sub	subset of data
y	observations

Other symbols

ℓ^2	Euclidean distance
\in	a member of
\mathbb{P}	probability
\mathbb{R}	real numbers

drawbacks, there has been increasing research towards the development of analytical or mathematical techniques to assist the calibration process [9].

For example, Sun et al. [45] proposed a pattern-based automated calibration approach that uses programmed logic to identify calibration parameters that would be tuned to minimize biases between simulated and actual energy consumption. In their “auto-tune” project, Chaudhary et al. [6] proposed a methodology that leverages on large databases of simulation results and an evolutionary (meta-heuristic optimization) algorithm to automate the calibration process. Optimization approaches involve defining an objective function such as minimizing the mean squared difference between simulation predictions and measured data. To prevent unreasonable parameter values, the objective function can be augmented with penalty functions that penalizes solutions that differ significantly from the base-case [5].

Due to its ability to naturally incorporate uncertainties, Bayesian calibration is another automated approach that is quickly gaining interest. In particular, the formulation proposed by Kennedy and O’Hagan [22] has been increasing applied to BEM [18,41,17,27,7,30] because it explicitly quantifies uncertainties in calibration parameters, discrepancies between model predictions and observed values, as well as observation errors. With the emergence of IoT and as more sensors get deployed in buildings, there is an opportunity to constantly update and adjust model parameters through continuous calibration. A Bayesian approach provides a flexible framework for dynamically updating a BEM. As new data arrive, the old data is not discarded but instead assimilated to the new data through the use of priors [3]. In other words, the previous posterior density acts as the prior for the current calibration, thus providing a very systematic framework for the continuous calibration or updating of the energy model.

Despite several successful applications of Bayesian calibration to BEM, challenges remain in its widespread adoption. First, Bayesian calibration is typically carried out using RWM or Gibbs sampling. An inherent inefficiency of these algorithms can be attributed to their random walk behavior as the MCMC simulation can take a long time zig-zagging while moving through the target distribution [12]. Second, current application as described in [18] is computationally prohibitive with larger datasets. As a result, Bayesian calibration has been limited to monthly calibration data.

The objective of this paper is to address these challenges by proposing a systematic framework for the application of Kennedy

and expertise of the modeler, making it harder to reproduce and thus restrict its widespread adoption. To overcome these

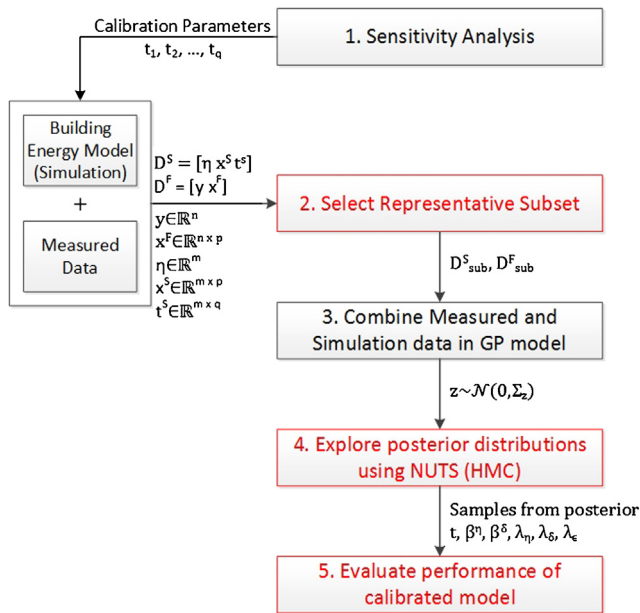


Fig. 1. Overview of the proposed Bayesian calibration framework.

and O’Hagan [22] formulation to BEM. The framework focuses on improvements to the current implementation of Bayesian calibration. In addition, the framework also includes evaluating the calibration process for both accuracy and convergence.

2. Method

2.1. Overview

Fig. 1 shows an overview of the proposed framework [8]. It is an extension of Kennedy and O’Hagan’s [22] Bayesian calibration framework that was first applied to BEM by Heo et al. [18]. The additions include using a representative subset of the measured data for the calibration and using the NUTS, an extension of the HMC for the MCMC sampling. In addition, for greater rigor in assessing the calibration process, the proposed framework includes evaluating both prediction accuracy (agreement between observations and calibrated predictions on test data) and convergence (multiple MCMC chains have converged to a common stationary distribution). The framework is summarized as follows with a detailed description of each step in the proceeding sections:

1. Use sensitivity analysis to identify influential parameters t that would be used for the calibration.
2. Select a representative subset D_{sub}^F from the field dataset D^F and a representative subset D_{sub}^S from the simulation dataset D^S .
3. Combine D_{sub}^F and D_{sub}^S in a GP emulator using the approach proposed by Higdon et al. [19].
4. Explore the posterior distributions of the calibration parameters t and GP hyper-parameters ($\beta^\eta, \beta^\delta, \lambda_\eta, \lambda_\delta$ and λ_ϵ) using the NUTS [21] for the MCMC sampling.
5. Evaluate performance of the calibrated model using two performance categories: (1) convergence of parameters of the posterior distribution, and (2) accuracy of the predictions by the calibrated model on a hold-out test dataset.

2.2. Sensitivity analysis

Sensitivity analysis is first used to identify parameters with the most influence over the model output. The Morris [32] method is

used to carry out the sensitivity analysis. This was implemented using the R sensitivity package [37]. The method of Morris belongs to the class of OAT design and is suitable when the number of input factors are so large that other variance-based approaches are computationally prohibitive [42], thus making it appropriate for use with building energy models [47,29,24]. This study uses the modified mean μ^* proposed by Campolongo et al. [4] and standard deviation σ as sensitive measures to help screen-out non-sensitive parameters.

2.3. Statistical formulation

In the proposed framework, Bayesian calibration was carried out following Kennedy and O’Hagan’s [22] formulation. This formulation explicitly models uncertainties in calibration parameters, discrepancies between model predictions and observed values, as well as observation errors (Eq. (1)).

$$y(x) = \eta(x, t^*) + \delta(x) + \epsilon(x) \tag{1}$$

Here, t^* represents the true but unknown values of the calibration parameters t , suggesting that the simulation model is a biased representation of the actual building system even in the ideal situation where $t = t^*$. Since the energy model $\eta(x, t)$ is a representation of the building’s actual performance, a discrepancy term $\delta(x)$ was included to correct any model inadequacy that could be revealed by the differences between the observed data and the model predictions.

Since the energy model could be computationally expensive to evaluate, a GP model is used as an emulator (surrogate) during the iterative calibration process. The GP model used to emulate the energy model was specified using a mean function that returns the zero vector and a covariance matrix given by Eq. (2) [19].

$$\Sigma_{\eta,ij} = \frac{1}{\lambda_\eta} \exp \left\{ - \sum_{k=1}^p \beta_k^\eta |x_{ik} - x_{jk}|^\alpha - \sum_{k'=1}^q \beta_{p+k'}^\eta |t_{ik'} - t_{jk'}|^\alpha \right\} \tag{2}$$

where λ_η is the variance hyperparameter; $\beta_1^\eta, \dots, \beta_{p+q}^\eta$ are the correlation hyperparameters; and α controls the smoothness of the function. $\alpha = 2$ (i.e., the ℓ^2 norm of their difference) was used, resulting in support for smooth functions. Similarly, the discrepancy term $\delta(x)$ was modeled using a GP model with a mean function that returns the zero vector and a covariance matrix of the form defined by Eq. (3) [19].

$$\Sigma_{\delta,ij} = \frac{1}{\lambda_\delta} \exp \left\{ - \sum_{k=1}^p \beta_k^\delta |x_{ik} - x_{jk}|^\alpha \right\} \tag{3}$$

where λ_δ and $\beta_1^\delta, \dots, \beta_p^\delta$ are the variance and correlation hyperparameters respectively. Similarly, $\alpha = 2$ was used for this GP model.

To establish a statistical relationship between the observed output $y(x^F)$ and the simulation output $\eta(x^S, t^S)$, the observed outputs $y_1, \dots, y_n \in \mathbb{R}$ are combined with the simulation outputs $\eta_1, \dots, \eta_m \in \mathbb{R}$ in a single $n + m$ vector $z = [y_1, \dots, y_n, \eta_1, \dots, \eta_m]$ [19]. The sampling model is then given by:

$$z \sim \mathcal{N}(0, \Sigma_z), \quad \Sigma_z = \Sigma_\eta + \begin{bmatrix} \Sigma_\delta + \Sigma_y & 0 \\ 0 & 0 \end{bmatrix} \tag{4}$$

where $\Sigma_\eta \in \mathbb{R}^{(n+m) \times (n+m)}$ (Eq. (2)); $\Sigma_\delta \in \mathbb{R}^{n \times n}$ (Eq. (3)); and $\Sigma_y = I_n / \lambda_\epsilon \in \mathbb{R}^{n \times n}$ is used to model observation errors. Therefore, the random variables in this model include the unknown calibration parameters t^F , the correlation hyperparameters ($\beta_1^\eta, \dots, \beta_{p+q}^\eta, \beta_1^\delta, \dots, \beta_p^\delta$), and the variance hyperparameters ($\lambda_\eta, \lambda_\delta$ and λ_ϵ). The non-linearity of the simulator makes it difficult to analytically sample from the joint posterior distribution. Therefore, a common approach is to use MCMC methods to explore

the posterior distribution of the calibration parameters and GP hyperparameters.

2.4. Limitations of Bayesian calibration

Although Kennedy and O'Hagan's [22] Bayesian calibration formulation explicitly models various sources of uncertainty, its application to BEM has been limited to low-dimensional monthly aggregated data (i.e., small number of calibration parameters and monthly calibration data). This is because when the output is at an hourly or daily resolution, the estimation of the covariance function becomes computationally prohibitive due to the resulting large sample sizes. Training GP models involve a runtime complexity of $\mathcal{O}(N^3)$, where $N=n+m$ is the number of samples the GP model is trained on. Based on previous studies [18,41,7,30], to learn about the calibration parameters t^F , each simulation is run at the same observed inputs x_1, \dots, x_n . Therefore, the resultant sample size m of the simulation data can grow very quickly as n increases. Second, RWM and Gibbs sampling are routinely used for Bayesian calibration. However, these algorithms may suffer from the "curse of dimensionality" and may take an unacceptable large number of iterations (longer runtimes) to achieve convergence to the high-dimensional posterior distribution $(t_1^F, \dots, t_q^F, \beta_1^\eta, \dots, \beta_{p+q}^\eta, \beta_1^\delta, \dots, \beta_p^\delta, \lambda_\eta, \lambda_\delta \text{ and } \lambda_\epsilon)$ [7].

To overcome the computational challenge of applying Bayesian calibration to high-dimensional datasets and large samples sizes, 2 strategies were employed:

1. Reducing the sample size by sampling a representative subset of the data for the calibration.
2. Using a more effective MCMC algorithm, NUTS [21] to generate samples from the posterior distribution. NUTS requires no manual tuning, is more efficient and converges more quickly in high-dimensional problems.

2.5. Representative subset selection

In the traditional design of experiments, the input factors x^F are systematically selected to help determine its relationship with the output of interest. However, in reality, energy modelers are not provided with the flexibility to configure the input factors x^F . Instead, they are typically provided with historical data containing measured values of x^F and $y(x^F)$ for the calibration of the building energy model. Therefore, the design space is the set of x^F values at which the building has been operated. This design space is defined by the field dataset $D^F = [y^F \ x^F]$, where $y^F = [y(x_1^F), \dots, y(x_n^F)]^T$ and $x \in \mathbb{R}^{n \times p}$ is the corresponding input factors. To learn about the unknown calibration parameters t^F , m simulations are then run at different combinations of (x^S, t^S) . Values of t^S for each simulation were determined using Maximin Latin hypercube sampling [44], which tries to cover as much parameter space as possible by maximizing the minimum distance between design points. Therefore, the corresponding design of experiments is the simulation dataset $D^S = [\eta^S \ x^S \ t^S]$, where η^S is the simulation predictions $\eta(x_1^S, t_1^S), \dots, \eta(x_m^S, t_m^S)$, $x^S \in \mathbb{R}^{m \times p}$ and $t^S \in \mathbb{R}^{m \times q}$.

As mentioned in the previous section, the resulting sample size $N=n+m$ that the GP model is trained on can increase very quickly as more observations n are used. Fortunately, having access to massive amounts of data does not imply that the calibration algorithm must be applied to the whole dataset. Sub-samples often provide the same accuracy with significantly lower computational cost [36]. The typical approach to selecting a sub-sample has been to manually select representative parts of the data (e.g. one week of summer data and one week of winter data) for the calibration and analysis. However, such a process requires domain knowledge and is harder

to be replicated and adopted by a non-expert. Therefore, the proposed approach uses random samples from the dataset. However, determining the correct sample size is often not intuitive. To overcome this, a statistical approach that is based on information theory is proposed.

Suppose there is a dataset D with r attributes and D_{sub} is its subset. To decide on the sample size, a metric known as sample quality [14] is used to measure how similar D_{sub} is to D (Eq. (5)). This metric makes use of the KL divergence [25] to measure the 'distance' of the selected subset D_{sub} from the whole dataset D .

$$Q(D_{sub}) = \exp(-J) \tag{5}$$

where

$$J = \frac{1}{r} \sum_{i=1}^r J_i(D_{sub}, D) \tag{6}$$

$$J_i(D_{sub}, D) = \sum_{k=1}^c (\mathbb{P}_k^{D_{sub}} - \mathbb{P}_k^D) \log \frac{\mathbb{P}_k^{D_{sub}}}{\mathbb{P}_k^D} \tag{7}$$

J is the averaged information divergence; $J_i(D_{sub}, D)$ denotes the Kullback–Leibler divergence (of the i th attribute) between the subset D_{sub} and the whole dataset D ; c is the number of bins or categories within the k th attribute; $\mathbb{P}_k^{D_{sub}}$ is the probability of occurrence for the k th value in D_{sub} ; and \mathbb{P}_k^D is the probability of occurrence for the k th value in D . In the calculation of $J_i(D_{sub}, D)$, each continuous attribute is first discretized. Since there may be zero entries in the sampled data D_{sub} , we smooth the distributions with a Dirichlet prior [16].

By definition, J is always larger than 0 [25,14]. Therefore $0 < Q \leq 1$, where $Q=1$ indicates no divergence between the sample and the entire dataset. From Eq. (6), it can be seen that the computation of the sample quality Q does not consider the joint information divergence but instead measures divergence between each attribute independently, which makes the calculation of the Q simpler and more straightforward.

Next, a sampling schedule is used to help determine the appropriate sample size. In this study, a geometric sampling schedule S is used (Eq. (8)).

$$S = \{s, a \cdot s, a^2 \cdot s, a^3 \cdot s, \dots\} \tag{8}$$

where $s > 0$ is the starting sample size; and $a > 1$ is the increment ratio. An example would be $\{10, 20, 40, 80, \dots\}$. Using the sampling schedule, the corresponding sample quality is computed and a curve depicting the relationship between sample size and sample quality is drawn (Fig. 3 and 9). From the curve, the appropriate sample size is selected taking into consideration both sample quality and computation cost. To ensure that the random samples used for the calibration has the same sample quality, the same subset that was used to draw the learning curve was also used for the calibration.

2.6. No-U-Turn-Sampler (NUTS) MCMC algorithm

Instead of the commonly used RWM [31] or Gibbs sampler [13], the proposed framework uses the NUTS [21], an extension to HMC algorithm for the MCMC sampling. The NUTS has been shown to be more effective for the Bayesian calibration of building energy models as compared to RWM and Gibbs sampler, achieving convergence to the posterior distribution more quickly and with significantly less iterations [7].

HMC is a variant of the MCMC algorithm that avoids the random walk behavior and sensitivity to correlated parameters that plague many MCMC methods, allowing faster convergence to high-dimensional posterior distributions [11,33,34,12]. To avoid the

random walk behavior, HMC borrows a concept from Hamiltonian dynamics, which describes an object's motion in terms of its position (location) and its momentum. In the context of applying HMC to the proposed Bayesian calibration framework, the location variables correspond to the parameters of the posterior distribution (i.e., the calibration parameters t^F and the GP hyperparameters $\beta_1^\eta, \dots, \beta_{p+q}^\eta, \beta_1^\delta, \dots, \beta_p^\delta, \lambda_\eta, \lambda_\delta$ and λ_ϵ). To make the algorithm move faster in the parameter space, HMC introduces an auxiliary momentum variable for each location variable. The goal is to use Hamiltonian dynamics to find a more efficient proposal or jumping distribution. However, a downside of HMC is that it involves multiple gradient evaluations per MCMC iteration, resulting in higher computation cost for each MCMC iteration as compared to RWM. Nonetheless, HMC is still more efficient since significantly less iterations is needed to achieve convergence and get a representative sampling [7,30].

The main part of HMC is the simultaneous update of the location and momentum variables. This update is carried out during each MCMC iterations and involves L leapfrog steps each scaled by a leapfrog step-size. Therefore, to run HMC, users need to provide values for (1) the leapfrog step-size (a scaling factor), and (2) the number of leapfrog steps L per iteration. This leads to time consuming initial runs required to manually tune both parameters. Poor choices of either parameter can result in an ineffective implementation of HMC [21,34]. To mitigate the challenges involved in tuning L , NUTS uses a recursive algorithm to automatically select the number of leapfrog steps L per iteration. It also automatically determines a value for the leapfrog step-size through a dual averaging scheme, thus making it possible to run HMC without requiring any user intervention. For this study, Bayesian inference with NUTS was implemented using R version 3.2.3 [38] with the package RStan [43].

2.7. Model evaluation

Two categories of performance metrics were used to assess the performance of the calibration process:

1. Assessing convergence of multiple MCMC chains to a common stationary distribution.
2. Validation of prediction accuracy on a hold-out test dataset that was not used for the calibration.

Trace plots of multiple MCMC chains and the Gelman–Rubin statistics (\hat{R}) were used to assess if convergence has been achieved [12]. Looking at multiple trace plots allow us to determine if the chains are well mixed and if different chains have converged to a common stationary distribution. Well-mixed chains indicate faster convergence and therefore faster computation. \hat{R} is the ratio of between-chain variance to within-chain variance and is based on the concept that if multiple chains have converged, there should be little variability between and within the chains [12]. For convergence, \hat{R} should be approximately 1 ± 0.1 .

CVRMSE and NMBE are used to evaluate the accuracy of the calibrated predictions. To prevent bias in the evaluation process, CVRMSE and NMBE are calculated on a hold-out test dataset that was not used for the calibration. CVRMSE provides a measure of how well the simulated data fits the actual values while NMBE serves as a good indicator of overall bias, providing an indication as to whether the predicted values tend to overestimate or underestimate the actual values. For hourly calibration data, CVRMSE of 30% and NMBE of 10% are considered acceptable according to ASHRAE Guideline 14 [1].

The use of two assessment metrics is in contrast with the current assessment methods, which accept a model as calibrated when CVRMSE falls below the threshold set by ASHRAE Guideline 14 [1,9].

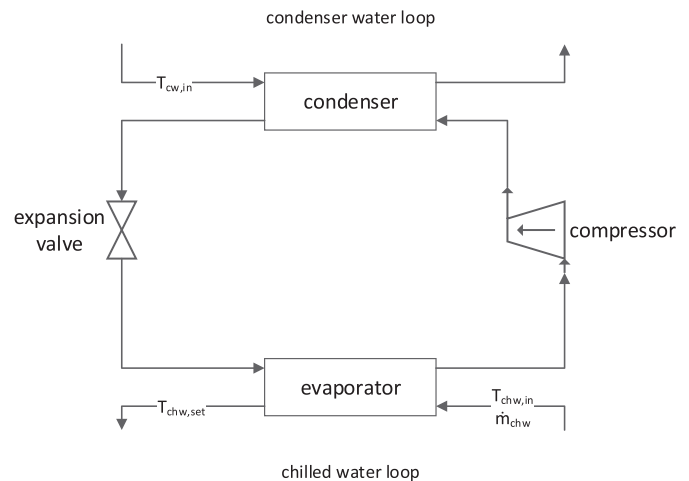


Fig. 2. Schematic of cooling process of TRNSYS type 666 water cooled chiller model.

Moreover, CVRMSE is typically calculated using data that was used to calibrate the model. Such a performance evaluation protocol is biased and may overfit the model, producing a calibrated model that performs poorly on unseen data. In addition, it is also important to check for convergence. If the MCMC algorithm has not proceeded long enough, the generated samples may grossly under-represent the target distributions [12], leading to misleading interpretation of the resulting posteriors.

3. Case studies

To demonstrate its application, the proposed Bayesian calibration framework was applied to two different case studies:

1. Case study 1: a TRNSYS model of a water-cooled chiller component for a mixed-use building located in a college in Singapore (Fig. 2).
2. Case study 2: an EnergyPlus model of the cooling system of a ten story office building located in Pennsylvania, U.S.A. (Fig. 7).

Table 1 provides a summary of the data used for the Bayesian calibration of both cases. The output was standardized to have zero mean and unit variance and the inputs normalized by scaling them in the range [0, 1]. This not only helps simplify the specification of the prior probabilities but also help achieve better estimates of the GP hyperparameters and ease maximum likelihood estimation [20,23]. Data preprocessing was carried out by removing instances containing missing, erroneous and inconsistent values. After preprocessing, the dataset for case studies 1 and 2 contained 1130 and 722 samples respectively. 30% of the data that was collected was randomly selected and withheld from the calibration to determine and validate the model's accuracy when predicting observations in the hold-out samples.

For both case studies, the following choices were made regarding the specification of the prior distributions:

- Calibration parameters t_1, \dots, t_q : independent uniform priors with range [0, 1] were specified for each t_1, \dots, t_q . This is consistent with the previous normalization, which scales them in the range [0, 1] (see Section 2.3).
- Correlation hyperparameters $\beta_1^\eta, \dots, \beta_{p+q}^\eta$: these correlation hyperparameters were reparameterized using $\rho_i^\eta = \exp(-\beta_i^\eta/4)$, $i = 1, \dots, p+q$ so that $0 < \rho_i^\eta < 1$ since $\beta_i^\eta > 0$ [20,15]. Independent Beta($a = 1, b = 0.5$) priors were specified for each ρ_i^η . Setting $a = 1$ and $0 < b < 1$ will put most of the prior support near 1, indicat-

Table 1
Summary of data used for the Bayesian calibration of the TRNSYS model and the EnergyPlus model.

Description	TRNSYS model (case study 1)	EnergyPlus model (case study 2)
Data collection period	1 Jan 2016 to 30 Apr 2016	1 Jun 2014 to 31 Aug 2014
No. of samples	Testing: 339 (30%) Training: 791 (70%)	Testing: 219 (30%) Training: 503 (70%)
Observed output $y(x^F)$ Simulation output $\eta(x^S, t^S)$	Measured energy consumption of water cooled chiller [kWh] Predicted energy consumption of water cooled chiller [kWh]	Measured energy consumption of the cooling system [kWh] Predicted energy consumption of the cooling system [kWh]
Input factors x	(a) x_1 : $T_{chw,in}$ [°C] (b) x_2 : \dot{m}_{chw} [kg/h] (c) x_3 : $T_{chw,set}$ [°C] (d) x_4 : $T_{cw,in}$ [°C]	(a) x_1 : Q_{load} [W] (b) x_2 : V_{frac} [–]
Calibration parameters t	(a) t_1 : Chiller rated capacity [kJ/h] (b) t_2 : Chiller rated COP [–]	(a) t_1 : Chiller 1 reference capacity [W] (b) t_2 : Chiller 1 reference COP [–] (c) t_3 : Chiller 2 reference capacity [W] (d) t_4 : Chiller 2 reference COP [–] (e) t_5 : Nominal capacity of cooling towers [W]
Priors before scaling to [0,1]	$t_1 \sim \mathcal{U}(1.80 \times 10^6, 2.69 \times 10^6)$ $t_2 \sim \mathcal{U}(5.6, 8.4)$	$t_1 \sim \mathcal{U}(5.23 \times 10^5, 7.84 \times 10^5)$ $t_2 \sim \mathcal{U}(5.5, 8.2)$ $t_3 \sim \mathcal{U}(1.95 \times 10^5, 2.93 \times 10^5)$ $t_4 \sim \mathcal{U}(1.9, 2.8)$ $t_5 \sim \mathcal{U}(4.40 \times 10^5, 6.60 \times 10^5)$
Posteriors (P2.5, P97.5)	t_1 : (2.01 × 10 ⁶ , 2.33 × 10 ⁶) t_2 : (7.1, 8.3)	t_1 : (5.26 × 10 ⁵ , 7.67 × 10 ⁵) t_2 : (5.7, 8.2) t_3 : (1.97 × 10 ⁵ , 2.78 × 10 ⁵) t_4 : (2.0, 2.8) t_5 : (4.47 × 10 ⁵ , 6.54 × 10 ⁵)
Field dataset	$D^F = [y \ x_1^F \ x_2^F \ x_3^F \ x_4^F]$ $D^F \in \mathbb{R}^{791 \times 4}$ $D_{sub}^F \in \mathbb{R}^{160 \times 4}$	$D^F = [y \ x_1^F \ x_2^F]$ $D^F \in \mathbb{R}^{719 \times 3}$ $D_{sub}^F \in \mathbb{R}^{80 \times 3}$
Simulation dataset	$D^S = [\eta \ x_1^S \ x_2^S \ x_3^S \ x_4^S \ t_1^S \ t_2^S]$ $D^S \in \mathbb{R}^{(791 \times 200) \times 7}$ $D_{sub}^S \in \mathbb{R}^{640 \times 7}$	$D^S = [\eta \ x_1^S \ x_2 \ t_1^S \ t_2^S \ t_3^S \ t_4^S \ t_5^S]$ $D^S \in \mathbb{R}^{(719 \times 200) \times 8}$ $D_{sub}^S \in \mathbb{R}^{640 \times 8}$

ing an expectation that only a subset of the inputs have an effect on the simulation output. Smaller values of b indicate an expectation that the output depends on a smaller number of inputs.

- Correlation hyperparameters $\beta_1^\delta, \dots, \beta_p^\delta$: similarly, these correlation hyperparameters were reparameterized using $\rho_i^\eta = \exp(-\beta_i^\eta/4)$, $i=1, \dots, p$. A more conservative independent $Beta(a=1, b=0.4)$ prior was assigned to each ρ_i^η because an even smaller subset of the inputs were expected to have an effect on the discrepancy term $\delta(x)$.
- Variance hyperparameter λ_η : a $Gamma(a=5, b=5)$ prior was assigned to this hyperparameter. Here, a represents the shape and b is the rate. Since the outputs were standardized to have unit variance, λ_η is expected to be close to one. Therefore, a Gamma prior with $a=b=5$ is suitable. In addition, this informative prior helps to stabilize the correlation hyperparameters [20,23]
- Variance hyperparameters λ_δ and λ_ϵ : $Gamma(a=1, b=0.0001)$ priors were specified for both λ_δ and λ_ϵ . This results in a prior that is quite uninformative. Therefore, if the data is uninformative about these parameters, the posterior would be large, which is consistent with very small discrepancy and observation errors respectively.

4. TRNSYS model (Case study 1)

In the first case study, the TRNSYS Type 666 model was used to model the water-cooled chiller (Fig. 2). This model relies on catalog data (provided in the form of lookup tables) to determine the chiller's performance at varying operating conditions [46]. At each time step, the model takes as inputs, (a) x_1 : chilled water inlet temperature $T_{chw,in}$; (b) x_2 : chilled water mass flow rate \dot{m}_{chw} ; (c) x_3 : chilled water setpoint temperature $T_{chw,set}$; and (d) x_4 : entering condenser water temperature $T_{cw,in}$. Combined with the model

parameters (chiller's rated capacity t_1 and COP t_2), the model is able to calculate the energy consumption of the chiller over a particular time period. Data collection for the model inputs (x_1, x_2, x_3 and x_4) and the chiller's energy consumption $y(x)$ took place between January 1 2016 and April 30 2016. Sensitivity analysis was not carried out for this case study because only two model parameters were identified as uncertain. Table 1 summarizes the inputs and output used for the calibration.

To learn about the calibration parameters, 200 TRNSYS simulations were run. Therefore, the experimental design corresponding to the field observations is a dataset $D^F = [y^F \ x_1^F \ x_2^F \ x_3^F \ x_4^F]$ where $D^F \in \mathbb{R}^{791 \times 5}$, and the experimental design for the simulations is a dataset $D^S = [\eta^S \ x_1^S \ x_2^S \ x_3^S \ x_4^S \ t_1^S \ t_2^S]$ where $D^S \in \mathbb{R}^{(200 \times 791) \times 7}$. Following Section 2.5, a representative subset D_{sub}^F and D_{sub}^S is sampled from the field dataset D^F and the simulation dataset D^S respectively. Considering both sample quality and computation cost, a sample size of 640 (D_{sub}^S) and 160 (D_{sub}^F) was used for the subsequent Bayesian calibration process (Fig. 3).

Using the formulation by Higdon et al. [19], the sampled field data D_{sub}^F and simulation data D_{sub}^S were combined using a GP model (see Section 2.3). NUTS (HMC) was then used to explore the posterior distributions. Based on a previous study [7], 500 iterations using NUTS was adequate to achieve convergence. Therefore, four independent chains of 500 iterations per chain were run. To be conservative, the first 250 iterations (50%) were discarded as warmup/burn-in to reduce the influence of the starting values.

4.1. Results for the TRNSYS chiller model

To ensure convergence to the stationary distribution, Fig. A12 shows the trace plots for all parameters of the posterior distributions. From the trace plots, it can be observed that the generated

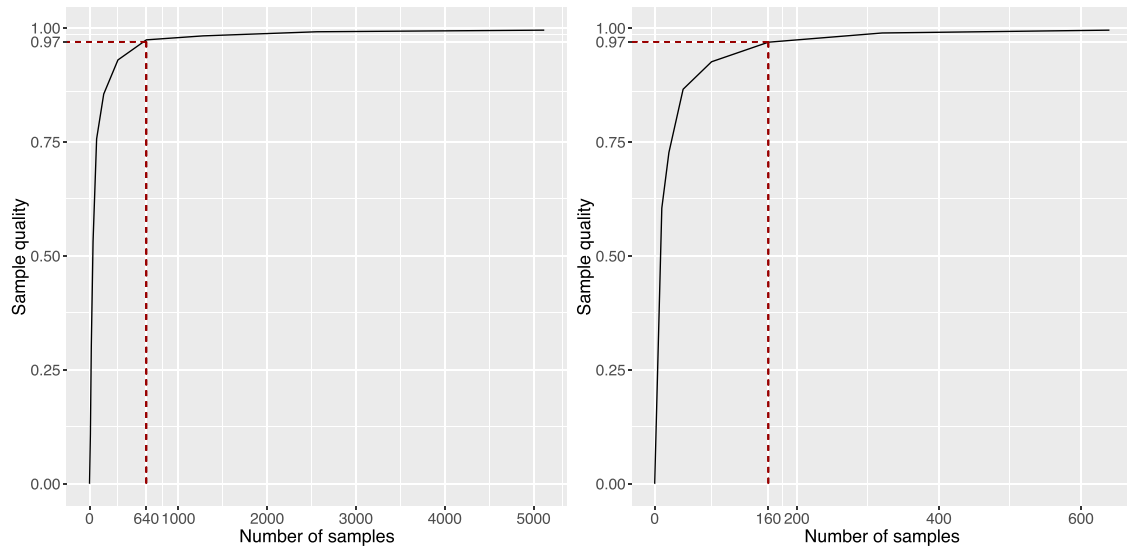


Fig. 3. Case study 1: number of samples against sample quality for different subsets of simulation data D^S (left plot) and field data D^F (right plot).

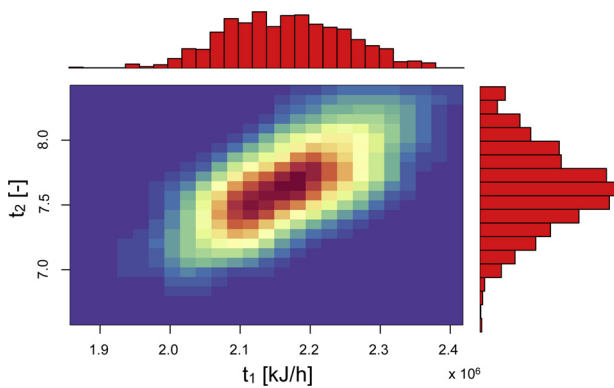


Fig. 4. Posterior distribution for the calibration parameters (chiller rated capacity t_1 and chiller rated COP t_2) with first 50% iterations discarded as burn-in.

samples are well-mixed and that all parameters of the posteriors ($t, \beta^\eta, \beta^\delta, \lambda_\eta, \lambda_\delta$ and λ_ϵ) have converged to a common stationary distribution. \hat{R} is also within 1 ± 0.1 for every parameter of the posterior.

The posterior distribution of the calibration parameters is shown in Fig. 4. Compared to its prior before scaling to $[0, 1]$ (Table 1), the posterior distribution for the chiller's capacity t_1 shows a reduction in variance and appears to be normally distributed with mean 2.17×10^6 kJ/h. Similarly, the posterior for the chiller's COP t_2 also shows a smaller variance compared to its prior (Table 1). Its posterior distribution also appears to be normally distributed with a 95% confidence interval of (7.1, 8.3), suggesting that the chiller is operating more efficiently than expected. Additionally, the 2 dimensional histogram in Fig. 4 shows a positive correlation between t_1 and t_2 . This is not surprising because the chiller's power consumption is directly proportional to the ratio of its capacity and its COP ($P_{chiller} \propto \frac{Capacity}{COP}$).

Given the focus on prediction accuracy, predictions from the calibrated model were evaluated using a hold-out test dataset. Fig. 5 compares the mean posterior predictions \hat{y} against the 339 hold-out observations. These are predictions by the calibrated model at input settings that were not part of the data used to train the GP model. Over the 339 hold-out samples, both CVRMSE (9.4%) and NMBE (-0.7%) were within the threshold for hourly calibration data [1]. The scatterplot in Fig. 5 also shows good agreement between the mean predictions and the actual values, with most

points lying close to the diagonal line $\hat{y} = y$. This is also illustrated by a large number of predictions falling within $\pm 1\sigma_y$ (right plot of Fig. 5). However, there are several predictions that have been over-estimated (up to $\pm 3\sigma_y$). Looking at these outliers reveal that they occur when the chiller energy consumption is low despite having high chilled water inlet temperature $T_{chw,in}$ and a low chilled water setpoint temperature $T_{chw,set}$ (right column of Fig. 5). A possible explanation is that the chiller is not operating as intended (i.e., the chilled water supply temperature is not meeting its setpoint temperature), and hence the lower energy consumption. Regardless, this suggests that the calibrated model is unable to capture this effect and that another input variable is probably needed to account for this discrepancy.

Fig. 6 shows the normalized mean posterior model discrepancy $\frac{\delta(x)}{y(x)}$ (middle column), the 95% confidence interval for the model predictions $\eta(x, t)$ (left column) and the 95% confidence interval for the calibrated predictions $\eta(x, t) + \delta(x)$ (right column). The predictions were computed using the hold-out measurements, which are also shown in Fig. 6. The figure shows that including the discrepancy term $\delta(x)$ increases the 95% confidence intervals of the predictions (i.e., the 95% confidence interval for $\eta(x, t) + \delta(x)$ is wider than without the discrepancy term), particularly at lower $T_{chw,in}$ and \dot{m}_{chw} . Including the discrepancy term $\delta(x)$ also places more observations within the 95% confidence interval of the predictions, indicating that the model is better calibrated with the discrepancy term. A reduction in CVRMSE (12.0–9.4%) using the mean posterior predictions was also observed when the discrepancy term was included. Mean posterior for $\frac{\delta(x)}{y(x)}$ also indicates that the calibrated model tends to slightly underestimate (positive $\frac{\delta(x)}{y(x)}$) the chiller's energy consumption at low $T_{chw,in}$, and slightly overestimate (negative $\frac{\delta(x)}{y(x)}$) the chiller's energy at high $T_{chw,in}$.

5. EnergyPlus model (Case study 2)

In the second case study, the cooling system of a ten-story office building was modeled using the following EnergyPlus objects (Fig. 7): (a) LoadProfile:Plant; (b) Chiller:Electric:ELR; (c) Pump:VariableSpeed; and (d) CoolingTower:SingleSpeed. Data collection took place between June 1, 2014 and August 31, 2014 and includes the cooling system's energy consumption, the chilled water flow rate, the chilled water supply and return temperatures, the condenser water flow rate, and the condenser supply and return water temperatures. Inputs to this model include [26]: (a) x_1 : the

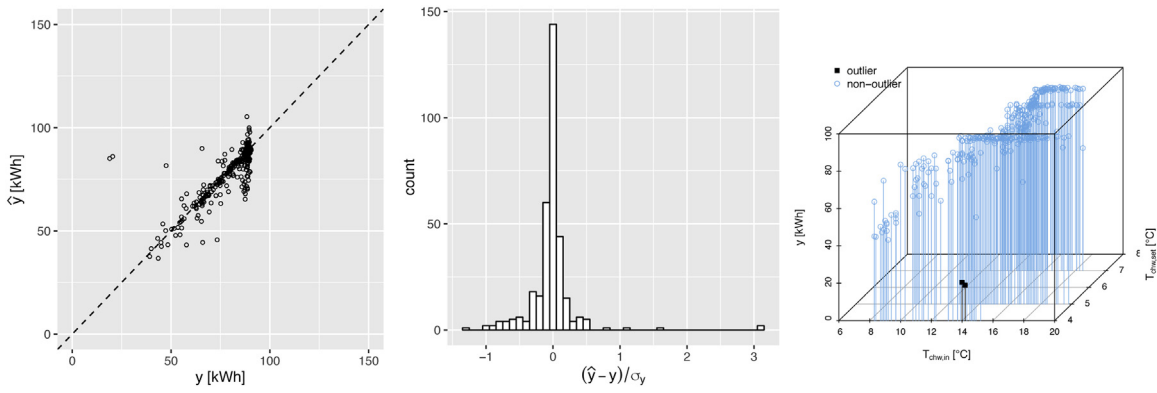


Fig. 5. Case study 1: actual observed values y against posterior mean predictions \hat{y} (left), histogram of residuals standardized by standard deviation of observed output (middle) and outliers based on residual histogram (right).

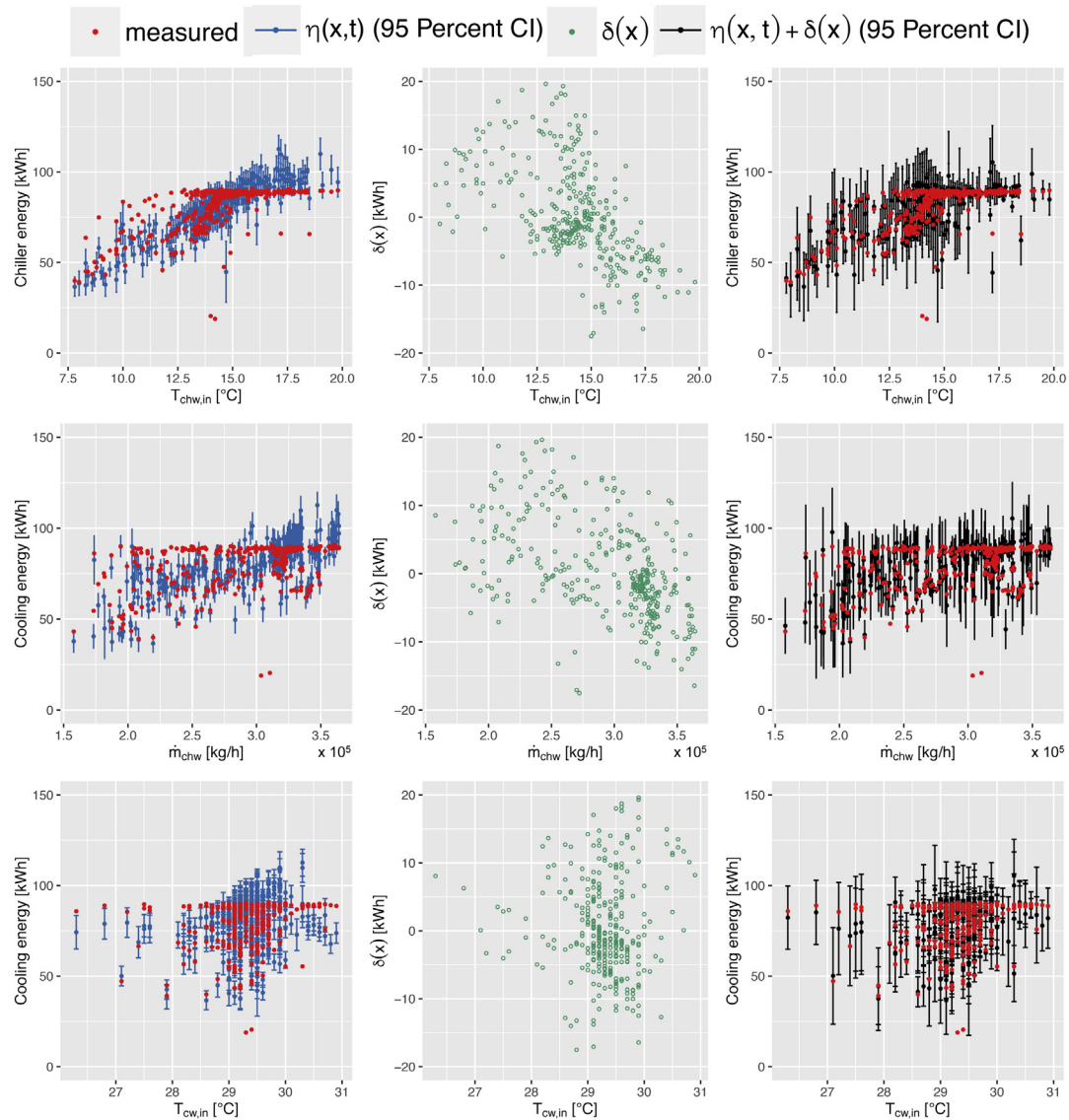


Fig. 6. Case study 1: 95% confidence interval for the calibrated simulation model (left column), the posterior mean for the model discrepancy $\delta(x)$ normalized by measured data $y(x)$ (middle column), and 95% confidence interval for the calibrated predictions (right column).

cooling coil load Q_{load} ; and (b) x_2 : the fraction of peak chilled water flow rate V_{frac} .

Ten model parameters $\theta_1, \dots, \theta_{10}$ were initially identified as uncertain. Table 2 shows their initial values and the range assigned

to each parameter. The initial values of each parameter were assigned based on measured data and as-built drawings and specifications. The design fan power θ_9 and nominal capacity θ_{10} of cooling towers 1 and 2 were modeled using a single random vari-

Table 2
Values used for sensitivity analysis in case study 2.

Model parameter	Symbol	Initial value	Min	Max
Chiller 1				
Reference capacity [W]	θ_1	653,378	522,702	784,053
Reference COP [-]	θ_2	6.86	5.49	8.23
Chiller 2				
Reference capacity [W]	θ_3	243,988	195,190	292,785
Reference COP [-]	θ_4	2.32	1.85	2.78
Chilled water pump				
Design power consumption [W]	θ_5	18,190	14,552	21,828
Motor efficiency [-]	θ_6	1.0	0.6	1.0
Condenser water pump				
Design power consumption [W]	θ_7	11,592	9274	13,911
Motor efficiency [-]	θ_8	1.0	0.6	1.0
Cooling towers 1 and 2				
Design fan power [W]	θ_9	11,592	9274	13,911
Nominal capacity [W]	θ_{10}	549,657	439,726	659,589

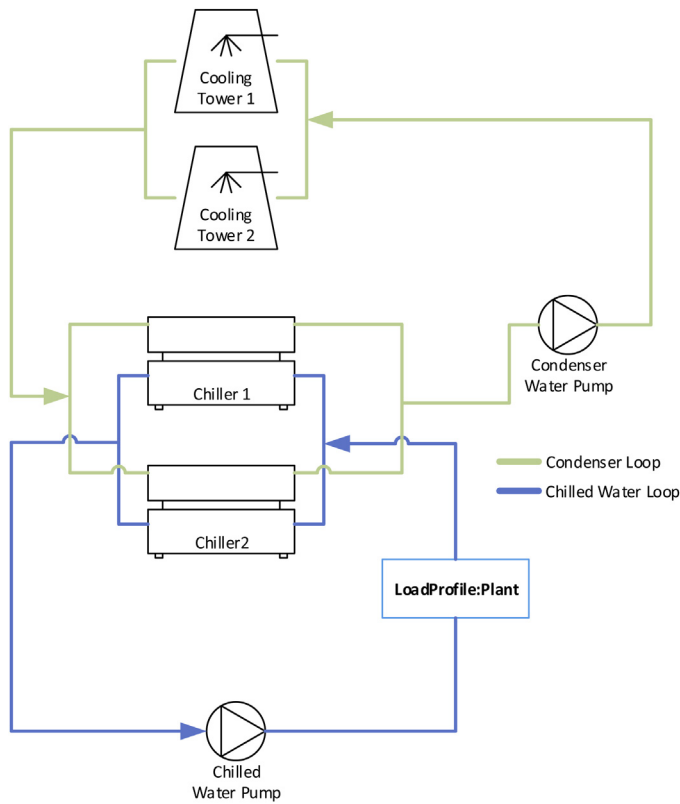


Fig. 7. Simplified representation of cooling system modeled using EnergyPlus.

able because both cooling towers have the same specifications and were installed at the same time. Since no prior information was available, pump motor efficiency was assigned a wide range of [0.6, 1.0]. To be conservative, the remaining parameters were varied $\pm 20\%$ of their initial values.

The result of the sensitivity analysis using the Morris [32] method is shown in Fig. 8. Based on the plot, only five parameters were found to have an influence on the model’s output, with the rest of the parameters having close to zero effect on the output. Here, t is used to denote the calibration parameters, i.e., the parameters in the set of uncertain parameters θ that have been identified as influential based on the sensitivity analysis. Table 1 provides a summary of the model inputs and output.

To learn about the calibration parameters, 200 EnergyPlus simulations were run. Therefore the experimental design corresponding to the field observations is a dataset $D^F = [y \ x_1^F \ x_2^F]$ where $D^F \in \mathbb{R}^{503 \times 3}$. Likewise, the experimental design for the simulations is a dataset $D^S = [\eta \ x_1^S \ x_2^S \ t_1^S \ t_2^S \ t_3^S \ t_4^S \ t_5^S]$ where $D^S \in \mathbb{R}^{(200 \times 503) \times 8}$.

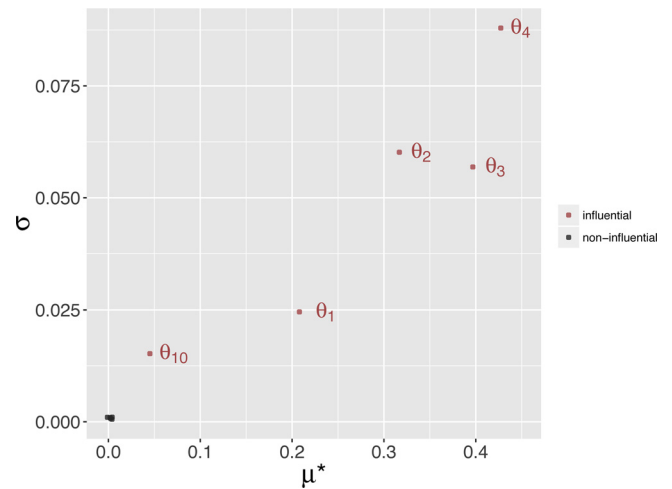


Fig. 8. Graphical plot of sensitive measures μ^* and σ for EnergyPlus model. The closer the parameters are to the upper right the more sensitive the parameter. Parameters close to the bottom left are non-influential parameters.

Similar to the first case study, a representative subset D_{sub}^F (80 samples) and D_{sub}^S (640 samples) is sampled from D^F and D^S (Fig. 9). The remaining samples from the field dataset D^F that were not used for the calibration were then used as a hold-out test dataset. The NUTS (HMC) was then used to explore the posterior distribution of the calibration parameters and GP hyperparameters. Four independent chains of 500 iterations per chain were run and the first 250 iterations (50%) were discarded as warmup/burn-in.

5.1. Results for the EnergyPlus model

The trace plots (Fig. A12) for parameters of the posterior distribution suggest that convergence has been achieved. This is similar to the trace plots for case study 1. Likewise, \hat{R} is also within 1 ± 0.1 for every parameter of the posterior. Next, the model’s prediction accuracy is evaluated using a hold-out test dataset. Over the 219 hold-out observations, both CVRMSE (6.0%) and NMBE (-0.3%) are within acceptable thresholds [1]. A scatterplot of actual values against the mean posterior predictions also shows that all but one point lie very close to the diagonal line of $\hat{y} = y$ (left column of Fig. 10). This is also illustrated in a histogram of standardized residuals (right column of Fig. 10), where it can be observed that the residuals are centered around zero and that all points are within $\pm 1\sigma_y$ with the exception of one point that was underestimated by about $1.6\sigma_y$. Looking at the measured data (right column of Fig. 10) reveals that the single point is an outlier and records a distinctly higher energy consumption as compared to other points

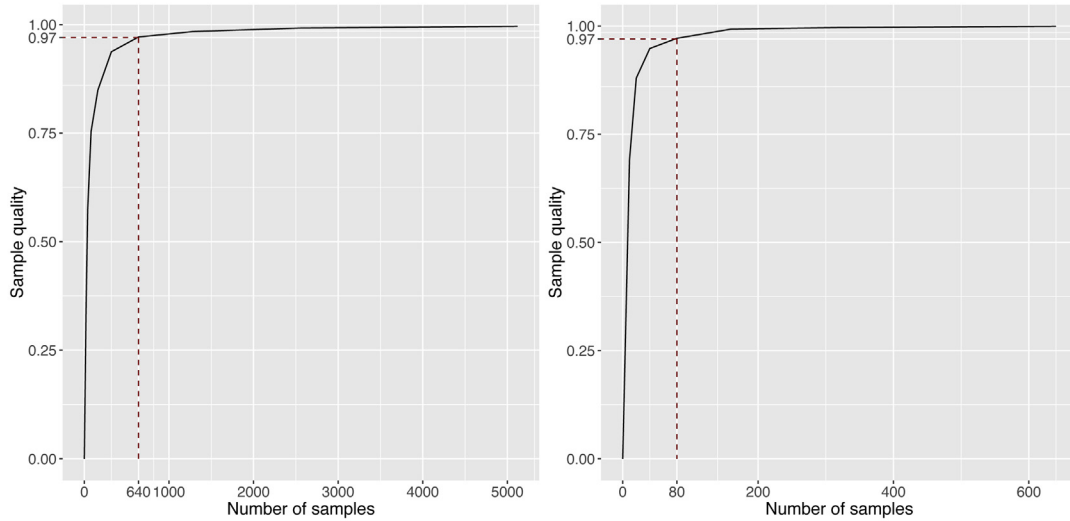


Fig. 9. Case study 2: number of samples against sample quality for different subsets of simulation data D^S (left plot) and field data D^F (right plot).

at the same load and flow rate, thus suggesting an error in measurement.

Fig. 11 shows the normalized mean posterior model discrepancy $\frac{\delta(x)}{y(x)}$ (middle column), the 95% confidence interval for the model predictions $\eta(x, t)$ (left column) and the 95% confidence interval for the calibrated predictions $\eta(x, t) + \delta(x)$ (right column). The figure shows that $\frac{\delta(x)}{y(x)}$ is negative across different values of Q_{load} and V_{frac} , suggesting that the model tends to overestimate the energy consumption of the cooling system. It increases to a peak when $Q_{load} \approx 5.5 \times 10^5$ W and gradually decreases as Q_{load} continues to increase. A notable reduction in overall uncertainty is visible with a significantly narrower 95% confidence interval when the discrepancy term is included in the calibrated predictions $\eta(x, t) + \delta(x)$ (third column as compared to first column of Fig. 11).

However, the magnitude of the discrepancy term is fairly large for this case study and does not stay constant over Q_{load} . Since the discrepancy term $\delta(x)$ is an indicator of how well the simulation output matches the observed values, the presence of a large discrepancy term makes interpreting the posterior distribution of the calibration parameters more difficult. The 95% confidence interval for the posterior distributions of the calibration parameters have ranges that are very similar to their prior distributions (Table 1), indicating that the calibration adds little to the uncer-

tainty of the calibration parameters t and suggesting that the data is non-informative about the calibration parameters.

6. Discussion

This study presented a systematic framework for the application of Kennedy and O'Hagen's [22] Bayesian calibration approach to BEM. The proposed framework focuses on improvements to the current implementation that was introduced by [18]. The improvements include:

- Reducing computation cost by:
 1. Selecting a representative subset of the entire dataset for the calibration.
 2. Using a more efficient MCMC algorithm, the No-U-Turn Sampler (NUTS) [21] that is an extension to Hamiltonian Monte Carlo (HMC) to explore parameters of the posterior distribution.
- Improving the rigor of assessing the performance of the calibrated process by:
 1. Assessing convergence of the calibration parameters and GP hyperparameters.

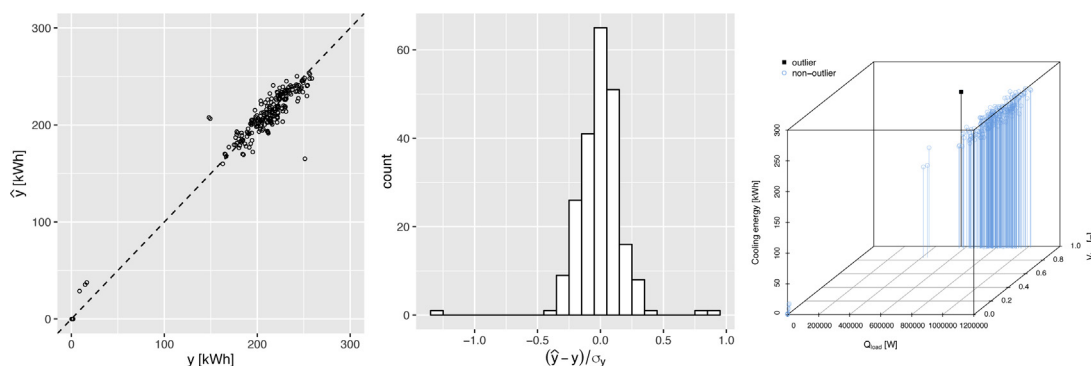


Fig. 10. Case study 2: actual observed values y against posterior mean predictions \hat{y} (left), histogram of residuals standardized by standard deviation of observed output (middle) and outliers based on residual histogram (right).

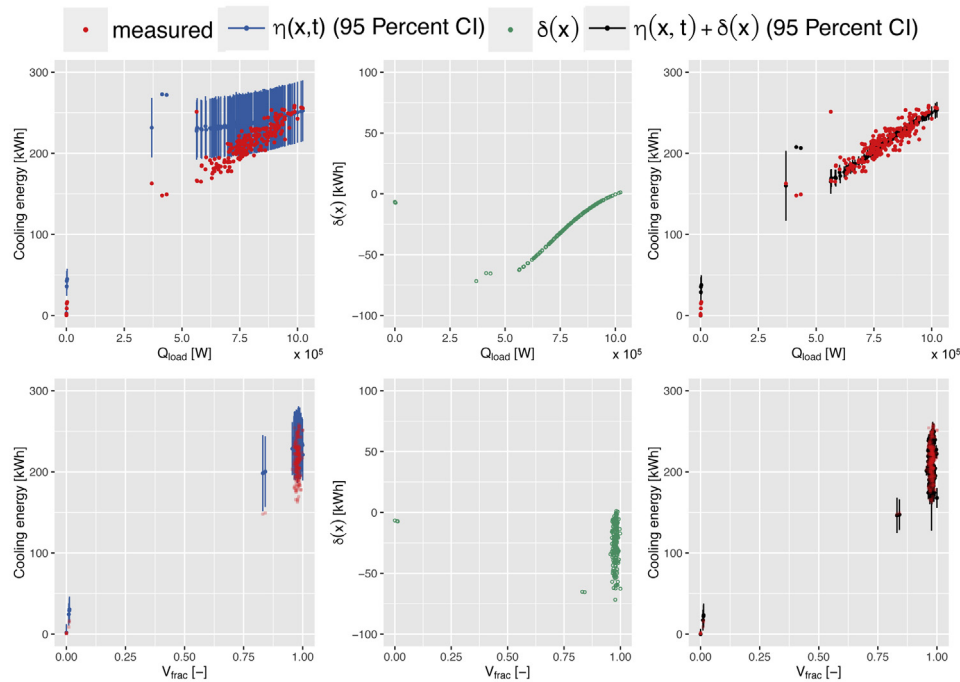


Fig. 11. Case study 2: 95% confidence interval for the calibrated simulation model (left column), the posterior mean for the model discrepancy $\delta(x)$ normalized by measured data $y(x)$ (middle column), and 95% confidence interval for the calibrated predictions (right column).

2. Validating the accuracy of the calibrated predictions against a hold-out test dataset.

Through the two case studies, this study demonstrated a Bayesian approach for calibrating building energy models against hourly data. In both case studies, visual plots (Figs. 5 and 10) showed good agreement between the calibrated predictions and measured test data, with most measurements being within the 95% confidence intervals of the predictions (Figs. 6 and 11). CVRMSE and NMBE for the hold-out test dataset were also less than 15% and 5% respectively, satisfying the thresholds set by ASHRAE guideline 14 [1]. This suggests that a small subset (<1000) of the entire dataset is adequate for the purpose of calibration, substantially reducing computation cost while still maintaining sufficient prediction accuracy. However, a limitation is that ad-hoc expert based manual selection may perform better than the random selection of the sampled subset, probably achieving similar accuracy with even lesser samples. On the contrary, the use of random samples would be more useful in an autonomous framework where automation of expert knowledge is difficult.

Based on the trace plots and \hat{R} values [12], NUTS was shown to be an effective MCMC algorithm for exploring and generating samples from the posterior distributions of the calibration parameters t , correlation hyperparameters (β^η and β^δ) and variance hyperparameters (λ_η , λ_δ and λ_ϵ). In both case studies, 500 iterations were sufficient to achieve adequate convergence (\hat{R} within ± 0.1 and trace plots of multiple chains indicates convergence to a common stationary distribution). In comparison, RWM and Gibbs sampling typically requires more than 10,000 iterations [7,8]. The low number of iterations can be attributed to good mixing as illustrated by the trace plots for each case study (Figs. A12 and A13), therefore making it suitable for the Bayesian calibration of BEM since the posterior is typically high-dimensional and can easily involve more than 10 dimensions.

Lastly, this study also investigated how the discrepancy term $\delta(x)$ affects the predictions by the calibrated model. In both cases, it was found that as intended, $\delta(x)$ was able to adjust for the varying

bias across different values of the input factors x , reducing overall bias and providing a better agreement between calibrated predictions and measured data. However, a low CVRMSE indicates a good fit between the resulting predictions and the observations, but does not justify that the posterior distributions of the calibration parameters are good estimates of its true value. How the discrepancy term affects the posterior distribution of the calibration parameters depends on the priors assigned as well as the system modeled. As mentioned previously, $\delta(x)$ was included to account for any model inadequacy that could be revealed by a discrepancy between the model predictions and the observed values. Therefore, when the magnitude of the discrepancy term is large, it makes interpreting the posterior distribution of the calibration parameters difficult and it is recommended that the model be investigated in greater detail if the purpose of the calibration is to provide realistic estimates of the calibration parameters. Although a large discrepancy is indicative of large model bias and suggests caution when interpreting the posterior distribution for the calibration parameters, the same is not true about its inverse. A small discrepancy does not suggest greater confidence in the posterior estimates for the calibration parameters. It is possible for models to have small discrepancies but be uninformative about the calibration parameters because the data did not contain any information about these parameters [8]. Therefore, it is recommended that both the posterior distributions and discrepancy terms be analyzed before interpreting the posterior for the calibration parameters.

Acknowledgements

The authors thank Dr. Kathrin Menberg and Dr. Yeonsook Heo for their technical assistance.

Appendix A. Trace plots

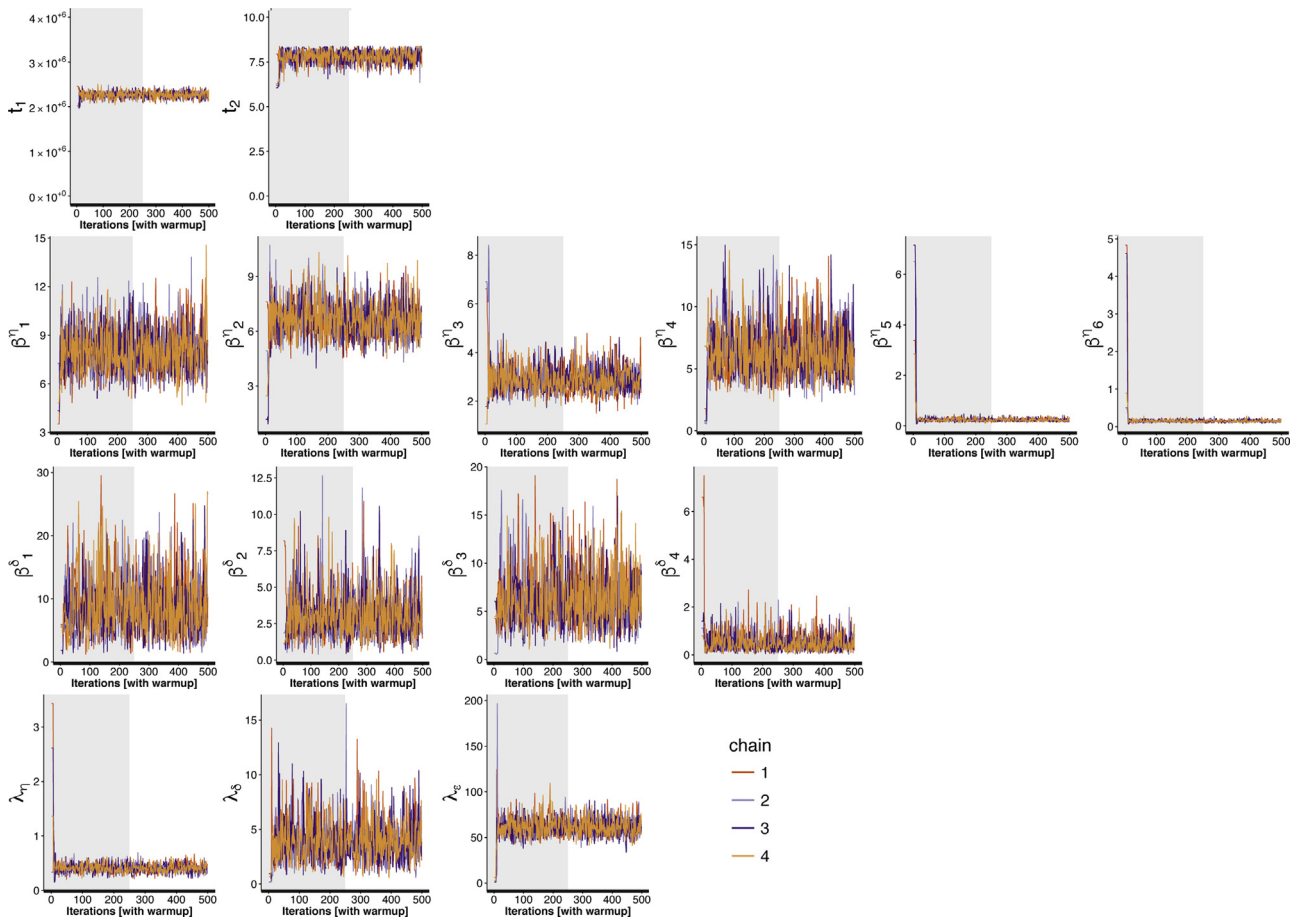


Fig. A12. Case study 1: trace plots of calibration parameters t , correlation hyperparameters (β^{η} and β^{δ}) and variance hyperparameters (λ_{η} , λ_{δ} and λ_{ϵ}).

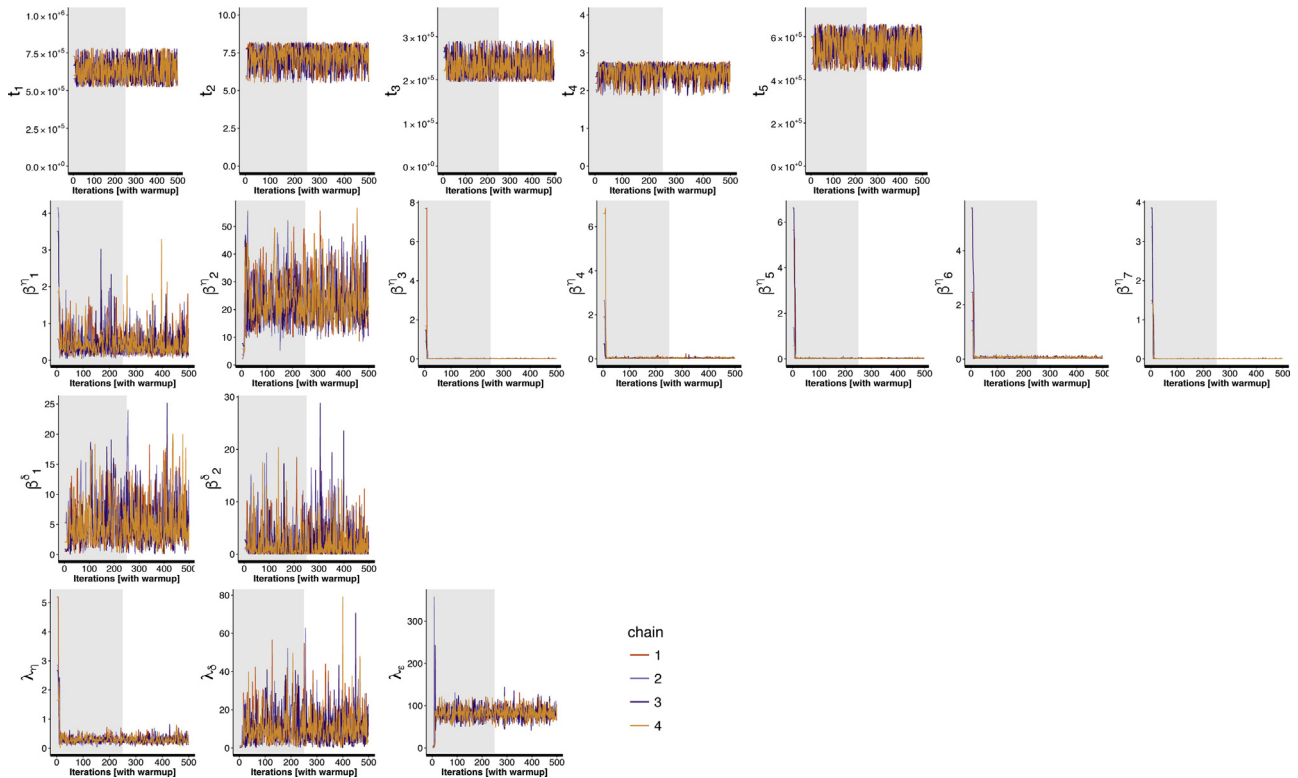


Fig. A13. Case study 2: trace plots of calibration parameters t , correlation hyperparameters (β^{η} and β^{δ}) and variance hyperparameters (λ_{η} , λ_{δ} and λ_{ϵ}).

References

- [1] ASHRAE, Guideline 14-2002, measurement of energy and demand savings, American Society of Heating, Ventilating, and Air Conditioning Engineers, Atlanta, Georgia, 2002.
- [2] K. Beven, A manifesto for the equifinality thesis, *J. Hydrol.* 320 (2006) 18–36.
- [3] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, K. Willcox, Y. Marzouk, Large-scale Inverse Problems and Quantification of Uncertainty, vol. 712, John Wiley & Sons, 2011.
- [4] F. Campolongo, J. Cariboni, A. Saltelli, An effective screening design for sensitivity analysis of large models, *Environ. Model. Softw.* 22 (2007) 1509–1518.
- [5] W. Carroll, R. Hitchcock, Tuning simulated building descriptions to match actual utility data: methods and implementation, *ASHRAE Trans.* 99 (1993) 928–934.
- [6] G. Chaudhary, J. New, J. Sanyal, P. Im, Z. O'Neill, V. Garg, Evaluation of “autotune” calibration against manual calibration of building energy models, *Appl. Energy* 182 (2016) 115–134.
- [7] A. Chong, K.P. Lam, A comparison of MCMC algorithms for the Bayesian calibration of building energy models, in: Proceedings of the 15th IBPSA Building Simulation Conference, 2017.
- [8] A. Chong, Bayesian calibration of building energy models for large datasets (Ph.D. thesis), Carnegie Mellon University, 2017.
- [9] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, *Renew. Sustain. Energy Rev.* 37 (2014) 123–141.
- [10] S. Duane, A.D. Kennedy, B.J. Pendleton, D. Roweth, Hybrid Monte Carlo, *Phys. Lett. B* 195 (1987) 216–222.
- [11] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, *Bayesian Data Analysis*, vol. 2, Taylor & Francis, 2014.
- [12] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* 6 (1984) 721–741.
- [13] B. Gu, B. Liu, F. Hu, H. Liu, Efficiently determining the starting sample size for progressive sampling, in: European Conference on Machine Learning, Springer, 2001, pp. 192–202.
- [14] S. Guillas, J. Rougier, A. Maute, A. Richmond, C. Linkletter, Bayesian calibration of the thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM), *Geosci. Model Dev.* 2 (2009) 137.
- [15] J. Hausser, K. Strimmer, Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks, *J. Mach. Learn. Res.* 10 (2009) 1469–1484.
- [16] Y. Heo, G. Augenbroe, D. Graziano, R.T. Muehleisen, L. Guzowski, Scalable methodology for large scale building energy improvement: relevance of calibration in model-based retrofit analysis, *Build. Environ.* 87 (2015) 342–350.
- [17] Y. Heo, R. Choudhary, G. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, *Energy Build.* 47 (2012) 550–560.
- [18] D. Higdon, M. Kennedy, J.C. Cavendish, J.A. Cafeo, R.D. Ryne, Combining field data and computer simulations for calibration and prediction, *SIAM J. Sci. Comput.* 26 (2004) 448–466.
- [19] D. Higdon, C. Nakhleh, J. Gattiker, B. Williams, A Bayesian calibration approach to the thermal problem, *Comput. Methods Appl. Mech. Eng.* 197 (2008) 2431–2441.
- [20] M.D. Hoffman, A. Gelman, The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.* 15 (2014) 1593–1623.
- [21] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 63 (2001) 425–464.
- [22] J.C. Kern, Bayesian process-convolution approaches to specifying spatial dependence structure (Ph.D. thesis), Duke University, 2000.
- [23] M.H. Kristensen, S. Petersen, Choosing the appropriate sensitivity analysis method for building energy model-based investigations, *Energy Build.* 130 (2016) 166–176.
- [24] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Stat.* 22 (1951) 79–86.
- [25] LBNL, Energyplus input output reference: the encyclopedic reference to EnergyPlus input and output, US Department of Energy, 2016.
- [26] Q. Li, G. Augenbroe, J. Brown, Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction, *Energy Build.* 124 (2016) 194–202.
- [27] G. Liu, M. Liu, A rapid calibration procedure and case study for simplified simulation models of commonly used HVAC systems, *Build. Environ.* 46 (2011) 409–420.
- [28] K. Menberg, Y. Heo, R. Choudhary, Sensitivity analysis methods for building energy models: comparing computational costs and extractable information, *Energy Build.* 133 (2016) 433–445.
- [29] K. Menberg, Y. Heo, R. Choudhary, Efficiency and reliability of Bayesian calibration of energy supply system models, in: Proceedings of the 15th IBPSA Building Simulation Conference, 2017.
- [30] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* 21 (1953) 1087–1092.
- [31] M.D. Morris, Factorial sampling plans for preliminary computational experiments, *Technometrics* 33 (1991) 161–174.
- [32] R.M. Neal, Probabilistic inference using Markov chain Monte Carlo methods, 1993.
- [33] R.M. Neal, MCMC using Hamiltonian dynamics, in: S. Brooks, A. Gelman, G.L. Jones, X.L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, vol. 2, Chapman and Hall/CRC, 2011, pp. 113–162 (Chapter 5).
- [34] A. Pedrini, F.S. Westphal, R. Lamberts, A methodology for building energy modelling and calibration in warm climates, *Build. Environ.* 37 (2002) 903–912.
- [35] F. Provost, D. Jensen, T. Oates, Efficient progressive sampling, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1999, pp. 23–32.
- [36] Pujol, G., Iooss, B., with contributions from Khalid Boumhaout, A.J., Veiga, S.D., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiot, L., Lemaitre, P., Ramos, B., Touati, T., Weber, F., 2016. sensitivity: Global Sensitivity Analysis of Model Outputs. <https://CRAN.R-project.org/package=sensitivity>. r package version 1.12.2.
- [37] R. Core Team, R: a language and environment for statistical computing, version 3.2.3, 2015 <https://www.R-project.org/>.
- [38] P. Raftery, M. Keane, A. Costa, Calibrating whole building energy models: detailed case study using hourly measured data, *Energy Build.* 43 (2011) 3666–3679.
- [39] T.A. Reddy, Literature review on calibration of building energy simulation programs: uses, problems, procedures, uncertainty, and tools, *ASHRAE Trans.* 112 (2006).
- [40] M. Riddle, R.T. Muehleisen, A guide to Bayesian calibration of building energy models, in: ASHRAE/IBPSA-USA Building Simulation Conference, 2014.
- [41] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, *Global Sensitivity Analysis: The Primer*, John Wiley & Sons, 2008.
- [42] Stan Development Team, Rstan: the r interface to stan, version 2.9.0, 2016 <http://mc-stan.org>.
- [43] M. Stein, Large sample properties of simulations using Latin hypercube sampling, *Technometrics* 29 (1987) 143–151.
- [44] K. Sun, T. Hong, S.C. Taylor-Lange, M.A. Piette, A pattern-based automated approach to building energy model calibration, *Appl. Energy* 165 (2016) 214–224.
- [45] J. Thornton, D. Bradley, T. McDowell, N. Blair, M. Duffy, N. LaHam, A. Naik, TessLibs 17: HVAC library mathematical reference, 2014.
- [46] W. Tian, A review of sensitivity analysis methods in building energy analysis, *Renew. Sustain. Energy Rev.* 20 (2013) 411–419.
- [47] F.S. Westphal, R. Lamberts, Building simulation calibration using sensitivity analysis, in: Proceedings of the 9th International IBPSA Conference, Citeseer, 2005, pp. 1331–1338.