ABSTRACT
Bayesian calibration provides substantial benefits over standard approaches to calibrating building energy models. It provides an automated process for optimizing parameter values that does not rely heavily on expert knowledge. It also accounts for multiple sources of uncertainty to generate an estimate for the range of possible outcomes rather than a single model prediction. However, using Bayesian calibration can be daunting. This paper aims to make it more accessible by outlining the key steps in setting up and using Bayesian calibration and by addressing design issues at each step. Key steps include quantifying uncertainty, screening parameters, conducting model runs to generate a metamodel, and performing Bayesian calibration based on the results of the model runs and measured field data. At each step, we explore the key decisions that must be made and the tradeoffs associated with using different approaches. We also discuss and compare software packages that can be used to aid in Bayesian calibration.

INTRODUCTION
Statement of the problem
Developing energy models for existing buildings is an essential part of evaluating many types of energy efficiency retrofit options (especially retrofits involving many building components). The models are used to develop energy use baselines for verifying energy savings from retrofits and to diagnose problems with building equipment or modes of operation. However, the development of energy models for existing buildings is fraught with uncertainty. There is uncertainty in most of the static building parameters (e.g., building geometry and building material properties), in the dynamic building parameters (e.g., occupancy, lighting, and plug loads), and in the stochastic building parameters (e.g., weather). To improve the predictive capability of these energy models, measurements of actual building energy use are compared to predictions, and model input parameters are adjusted to reduce the difference between the two. This procedure is known as model calibration. While expert practitioners can often significantly reduce the model/measurement discrepancy, such calibration is subject to the biases of the modeler. Different modelers can calibrate different parameters without necessarily considering the relative uncertainty of the parameters or the sensitivity of the model to adjustment. Further, even after the calibration, the uncertainty of the calibrated parameters and that of the overall model prediction are still unknown. As a result, often the confidence in the predictions of the calibrated energy models is lacking. Even worse, models can be incorrectly calibrated, and then the error in the predictions of the energy savings from retrofits can be significant.

A solution: Bayesian calibration
The Bayesian calibration of building energy models can help address many challenges faced by practitioners. In deciding which parameters to adjust to calibrate a model, the ability to produce a good fit to the data must be balanced against knowledge about the uncertainty of each parameter estimate in order to determine the values that are most likely to represent reality. Bayesian methods are ideal for incorporating both these pieces of information into the calibration process in a systematic way. Knowledge about parameter uncertainty is incorporated by specifying a prior distribution for each parameter, which generally includes a most likely value and a range of possible values. Model runs are used to determine which sets of parameters are most likely to generate the observed data. Bayes’s theorem is then used to calculate a posterior parameter distribution, \( p(\theta \mid y) \), which gives the sets of parameters \( \theta \) that are most consistent with both the prior distributions, \( p(\theta) \), and the observed data, \( y \), given the likelihood function \( p(y \mid \theta) \).
Bayesian calibration of EnergyPlus model runs on virtual buildings with known true values for the calibration parameters. Under a range of different audit levels, Bayesian calibration was found to substantially reduce the parameter uncertainty and move the parameter estimates closer to their true values.

Limitations and challenges
Bayesian calibration does not solve all the challenges associated with calibrating building energy models. If there is not enough energy data to distinguish between different possible explanations for the data that are observed, then the Bayesian calibration process may generate wide ranges for the parameter distributions, limiting the model’s predictive ability. However, this is still an improvement over alternative approaches that are similarly uncertain but do not explicitly show this uncertainty, thereby possibly generating unwarranted confidence in the results.

Computation times can also put some limits on the application of Bayesian calibration, because run times increase quickly with the addition of more output data or calibration parameters. As a result, calibration against detailed data, such as hourly electricity use data, may not be feasible using standard techniques. Transformations can be used to reduce the dimensionality of the output to decrease calculation times (Higdon 2008; Wilkinson 2011), but more work is needed to explore the application of these methods to the calibration of building energy models.

Perhaps the most important factor limiting the adoption of Bayesian calibration for use with building models is the learning curve associated with the use of new calibration techniques. Bayesian calibration is conceptually different from standard approaches. It can be difficult to understand how it works and even more challenging to implement it.

A goal of this paper is to make Bayesian calibration more accessible by presenting an overview of the steps needed to perform Bayesian calibration, alternative design options, and illustrative examples. Some of the software packages that could be used are briefly reviewed. This is followed by a discussion of potential applications and a review of the results from a validation study that demonstrates the utility of Bayesian calibration for use with building models.

GUIDE TO BAYESIAN CALIBRATION

Uncertainty quantification
The first step in the Bayesian calibration process is to quantify the uncertainty in model input parameters in the form of a probability density function. For parameters that originate from measurements of the building, the measurement standard describing the measurement (if there was such a standard) will sometimes describe the uncertainty of the results. If such measurement standards were not used or if parameters were estimated from qualitative descriptions of building construction, one can turn to the literature reviews of parameter uncertainty (Macdonald 2002; Knight and Dunn 2003; Persily and Gorfain 2004; SCE 2005; de Wit 2001; ASHRAE 2009, 2011). This step is one of the most time consuming in the entire calibration process and would be improved with the development of a standardized database of uncertainty information.

When only a range of parameter values can be determined, a uniform probability distribution between the ranges should be assigned to the input. When a likely value can be determined in addition to the range, a triangle distribution or beta distribution peaking at the.
likely value and bounded by the range should be assigned. The exact form of the initial distribution is not particularly important, as long as the uncertainty range and general shape characterize the uncertainty of the parameter well. However, if the assessment of the initial distribution is accurate, the method will converge to an accurate posterior distribution faster than if the prior distribution is inaccurate. More importantly, if the prior distribution significantly underestimates the actual uncertainty of the parameter, the calibration process may fail.

Parameter screening

The next step is to determine a subset of the parameters on which to focus in subsequent calibration steps. It is needed to keep run times manageable. The parameters selected are those that have the most influence on model results when their values are adjusted within specified uncertainty ranges. Highly uncertain parameters (such as outdoor airflow rates) will be adjusted more and are therefore likely to have more influence on model results than are less uncertain parameters (such as insulation conductivity). Tests in a controlled virtual-reality setting have shown that little is gained by including additional parameters beyond the most important three to six (Heo et al. 2014).

A number of methods are possible for parameter screening. Here, we give an overview of two of the most prominent methods: the Morris method and sequential bifurcation. The Morris method is designed to evaluate the impact of changing one parameter at a time while covering a range of the parameter space (Morris 1991; Campolongo et al. 2005). The average impact of changing each parameter on model outcomes is computed. These estimates can then be used to select a set of parameters with the greatest impacts.

Sequential bifurcation is an alternative technique for parameter screening that can substantially reduce the number of model runs needed, especially when the number of parameters being tested is large (Bettonvil and Kleijnen 1997; Wan et al. 2010). This is done by iteratively dividing the parameters into groups, testing the effect of changing multiple parameters together, and eliminating them if the combined effect is below a threshold. This approach is designed for monotonic relationships between parameters and model outputs for which the direction of these relationships is known. Parameters for which the direction is unknown can be handled separately (Bettonvil and Kleijnen 1997).

Experiment design for generating a metamodel

Once a set of parameters has been selected for calibration, the next step is to design a simulation experiment and to run the model at the input values specified by the design. In addition to the parameters selected for calibration, the experiment design may also include observable inputs, such as weather data or control variables. The results of these model runs are then used to generate a metamodel, or emulator, that is used in Bayesian calibration. The experiment design plays an important role in ensuring that the metamodel can predict model results as well as possible over the full parameter space being evaluated.

A common approach for this step is to use Latin hypercube sampling (LHS). A random Latin hypercube sample of size n is one in which each parameter’s values are independently distributed between n evenly sized intervals in the parameter’s range. This design ensures that (1) a full range of values for the parameters with the highest impact are explored, and (2) model runs that vary only the values of the parameters with less impact are avoided. It also generally leads to a better covering of the full parameter space than does random Monte Carlo sampling.

A variation on LHS that has been used with Bayesian calibration is orthogonal-array-based LHS (Tang 1993). This design maintains the properties of LHS but ensures even better coverage of the full parameter space than does a random LHS. Other possible approaches include low discrepancy sequences (Bratley et al. 1992; Kocis & Whiten 1997) and sparse grids (Bungartz and Griebel 2004).

Calibration

Once the parameters have been chosen and the model has been run at a range of parameter values, Bayesian calibration can be used to generate distributions for the selected parameters. The inputs to Bayesian calibration consist of a set of observed data and a set of data generated by the model runs. The observed data include model outputs (e.g., utility data) and observable inputs (e.g., weather data). The model run data set includes model outputs, observable inputs, and parameter values for each model run. The goal of the calibration process is to generate distributions of possible values for the true parameters that are as consistent as possible with the observed data, with the model runs, and with prior information about the parameters.

Kennedy and O’Hagan (2001) originally formulated a mathematical framework for Bayesian calibration, and their framework has been the basis for nearly all subsequent work in this area. Posterior distributions for the parameters are generated from a combination of prior knowledge about the parameters and a calculation of the likelihood that a set of parameters would produce the observed data. The likelihood calculation is based on a model of how the observed data, y, relates to the calibration parameters, \( \theta \), and the observable inputs, \( x \).
A general relation among calibration parameters, observable inputs, and observed data is written in terms of a model prediction, $\eta(x, \theta)$, model inadequacy, $\delta(x)$, and measurement error $\epsilon(x)$ as

$$y(x) = \eta(x, \theta) + \delta(x) + \epsilon(x). \quad (2)$$

Both the model prediction term $\eta(x, \theta)$ and the model inadequacy term $\delta(x)$ are modeled as Gaussian processes—a flexible class of stochastic functions that does not impose a shape of the function, except that the joint distribution of the function’s output at any finite set of inputs must be multivariate normal. The properties of the Gaussian process are defined by its mean and covariance functions. The covariance function is of particular interest because it determines the smoothness properties of the Gaussian process. The form used in Kennedy and O’Hagan (2001) and most subsequent work is:

$$c(x, x') = \sigma^2 \exp\left(-\sum_{j=1}^{q} \omega_j (x_j - x'_j)^2\right) \quad (3)$$

The smoothness of the Gaussian process results from the high correlation between output distributions when the input values are close together. The hyperparameters $\sigma$ and $\omega_j$ determine how quickly output values can be expected to change as the input parameter changes.

A simple Gaussian process regression example is useful for illustrating how Gaussian processes are used and the role that hyperparameters play. Figure 1 shows the results from fitting three data points with a Gaussian process with the covariance function described in equation (3), hyperparameters $\omega_j = 1$ and two values (5 and 10) for $\sigma^2$. The estimated amount of uncertainty between the data points varies significantly depending on the value of $\sigma^2$.

\[\text{Figure 1 Gaussian process regression with different hyperparameters, } \sigma^2\]

In this example, the hyperparameter values were fixed. A better approach is to estimate the most likely values for these parameters by using Bayesian methods. This can be done by including hyperparameters as well as model parameters in the calibration process. Hyperparameters and parameters can be calibrated either simultaneously or through a two-stage process in which some hyperparameters are selected in advance (Kennedy and O’Hagan 2001; Bayarri et al. 2007).

Joint posterior distributions for the parameters and hyperparameters are calculated by multiplying the prior distributions by the likelihood, as in equation (1). The likelihood of generating the observed output, given a specified set of parameter and hyperparameter values, is found by looking at the joint multivariate normal distribution for the full set of output values from both the model runs and the observed data. The mean and covariance matrices for this distribution can be calculated for each possible set of parameters and hyperparameters being tested by using those parameter values in equation (2) and by using the mean and covariance functions for the Gaussian processes $\eta(\cdot)$ and $\delta(\cdot)$. It is then possible to calculate the likelihood of pulling the observed set of outputs from this multivariate normal distribution. For details of this calculation, see Kennedy and O’Hagan (2001).

The denominator of the posterior distribution cannot be calculated analytically, so it is easiest to draw a sample from the distribution by using a Markov chain Monte Carlo (MCMC) process, with the Metropolis-Hastings acceptance criteria. During the MCMC process, the parameter space is explored, and at each new point in the parameter space, the relative posterior density of the old and new point are compared to determine whether the new point is accepted. The end result of this process is a random sample from the posterior distribution of the parameter space, which can be used to generate model predictions.

A simple illustrative example with only one parameter (infiltration rate) is useful to understand how Bayesian calibration works. Figure 2 shows measured monthly energy use over 12 months and model run results for four values of the infiltration rate. It shows that a higher infiltration rate leads to increased energy use during the winter months. The ‘observed’ values were generated from model runs based on an assumed ‘true’ infiltration rate of 4.38 m$^3$/h/m$^2$ (infiltration volume flow rate per unit envelope area), with some random error added. Figure 3 shows the resulting posterior distribution for the infiltration parameter, along with its prior distribution and true value. The posterior distribution is centered at the true value, with a range that indicates how much uncertainty there is. The uncertainty is much

© 2014 ASHRAE (www.ashrae.org). For personal use only. Reproduction, distribution, or transmission in either print or digital form is not permitted without ASHRAE’s prior written permission.
less than that of the prior distribution, but it could be reduced more with more model runs close to the true parameter value. On the other hand, if the model did not match the data as closely, we would expect to see more uncertainty, since the calibration process would determine that the model inadequacy and measurement error terms were greater.

With more than one parameter, the calibration process generates a joint distribution for the calibrated parameters. Figure 4 shows an example of a joint distribution that results from calibrating three parameters. The diagonals show the distributions of individual parameters, and the off-diagonals show pairwise joint distributions. The negative correlation between the outdoor airflow and the infiltration rate reflects the fact that when the outdoor airflow is higher, a lower value for the infiltration rate is needed to match the observed energy data, and vice versa. The reverse is true of the relationship between outdoor airflow and heating system efficiency: An increase in outdoor airflow must be offset by an increase in heating system efficiency to achieve the same energy performance.

**Prediction**

Once calibration is completed, the posterior distributions for the parameter values can be used to predict model results or to predict the impacts of a proposed retrofit. The results that are generated will provide a full distribution of possible outcomes. Predictions can be generated either by re-running the full model using parameter values sampled from the posterior distribution or by using the meta-model that was generated for the calibration process. Re-running the full model can increase the precision of the estimates, but it requires more computational effort. Figure 5 shows an example from Heo et al. (2014) of the type of results that are generated, with a mean and confidence interval for each model output. It also shows the improved accuracy and precision of predicted model results over those of the uncalibrated model.

**Software packages**

A number of software packages that have been developed can aid in the Bayesian calibration process, so new users do not have to build the code from scratch. For parameter screening and designing experiments, there are a number of optionse, including the R package ‘sensitivity,’ the DAKOTA project, the Python library SALib, and the Eikos sensitivity analysis toolbox. The options for the core Bayesian calibration step are more limited. We know of at least three publicly available packages that can be used to generate posterior parameter distributions and model predictions, given a set of model run results and observed data: two R packages, SAVE and BACCO, and a Matlab package, GPM/SA. However, there are challenges associated
Overview of potential applications

The Bayesian calibration method can be applied to virtually any type of building energy model for which input parameters are specified, some sort of outputs are generated, and measured values of those outputs exist as a function of some variation on the inputs. Typically the measured data used for calibration are the values for monthly overall electricity or gas use from utility bills, but data from other time intervals (e.g., daily or hourly) or data specific to particular equipment or areas (e.g., chiller electrical input power, monthly average room temperature) could be used.

We believe this calibration method will be particularly valuable when model parameters have high levels of uncertainty. This is often the case before a detailed building audit has been performed. In such cases, the variable selection methods and subsequent Bayesian calibration can systematically improve confidence in retrofit savings predictions enough so that accurate energy efficiency retrofit choices can be made without the need for a detailed audit. When the calibration of many input parameters is desired and when highly detailed measurements of the building are available, machine learning methods (New et al. 2012) may be a more appropriate approach to calibration.

Validation results

Although the application of Bayesian calibration to building energy models is still in its early stages, the results found so far have been promising. Studies by Heo et al. (2012, 2014) have demonstrated the effectiveness of using Bayesian calibration with different energy models, types of buildings, and information levels. We focus here on Heo et al. (2014), which used a simulated virtual reality setting to test the effectiveness of calibration in a controlled environment where ‘true’ values are known. The observed data were generated in EnergyPlus model runs based on a U.S. Department of Energy (DOE) commercial reference building; specifically a medium-sized office building built before 1980 in Chicago (DOE 2012). A range of parameter values randomly sampled from the prior distributions were input into EnergyPlus along with data on the average weather in Chicago. A building was selected from the 95th percentile of the resulting distribution as measured by total energy use. The 95th percentile was used so that it would test the effectiveness of calibration when parameters differed significantly from their prior distributions. A small amount of measurement error was then added to the EnergyPlus model output to generate the ‘observed’ utility bill that was used for calibration.

Heo et al. (2014) tested the effectiveness of Bayesian calibration by using three uncertainty ranges for the parameters based on different audit levels. Figure 6 shows the results of the calibration process for one representative parameter: heating system efficiency with a Level 3 audit. The posterior distribution is closer to the ‘true’ parameter value than to the prior distribution. The results for all parameters and audit levels are shown in Heo et al. (2014).

To measure how close a distribution is to the true value, the root mean squared error (RMSE) and the coefficient of variation of RMSE (CVRMSE) are used. Table 1 shows the improvement in the CVRMSE for each of the four parameters and three audit levels explored. In all cases, the CVRMSE is reduced, and in some cases, it is reduced substantially.

![Figure 5 Comparison of energy use predicted by calibrated models (red circles), uncalibrated models (blue squares), and utility bills (black stars) (Source: Heo et al. 2014)](image-url)
Table 1 CVRSME values comparing prior distributions (UNCAL) and posterior distributions (CAL) with ‘real’ values (Source: Heo et al. 2014)

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>UNCAL</th>
<th>CAL</th>
<th>DIFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit Level 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infiltration rate</td>
<td>0.49</td>
<td>0.20</td>
<td>0.29</td>
</tr>
<tr>
<td>Heating sys. eff.</td>
<td>0.15</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>Infiltration reduc.</td>
<td>0.13</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>Heating setpoint T</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Audit Level 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outdoor airflow</td>
<td>0.38</td>
<td>0.30</td>
<td>0.08</td>
</tr>
<tr>
<td>Infiltration rate</td>
<td>0.48</td>
<td>0.39</td>
<td>0.09</td>
</tr>
<tr>
<td>Heating sys. eff.</td>
<td>0.12</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>Appliance power</td>
<td>0.27</td>
<td>0.22</td>
<td>0.05</td>
</tr>
<tr>
<td>Audit Level 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outdoor airflow</td>
<td>0.47</td>
<td>0.11</td>
<td>0.36</td>
</tr>
<tr>
<td>Infiltration rate</td>
<td>0.46</td>
<td>0.30</td>
<td>0.16</td>
</tr>
<tr>
<td>Heating sys. eff.</td>
<td>0.13</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Appliance power</td>
<td>0.31</td>
<td>0.14</td>
<td>0.17</td>
</tr>
</tbody>
</table>

The RMSE and CVRMSE for the model predictions are reduced even more in Table 2.

Table 2 Evaluation of model predictions against utility bills (Source: Heo et al. 2014)

<table>
<thead>
<tr>
<th>AUDIT</th>
<th>UNCAL</th>
<th>CAL</th>
<th>RMSE</th>
<th>CVRMSE</th>
<th>RMSE</th>
<th>CVRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3</td>
<td>1.3×10^{11}</td>
<td>0.29</td>
<td>0.2×10^{11}</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CONCLUSION

Bayesian calibration can be an effective tool for calibrating building energy models. It provides a way of automatically calibrating to observed energy use data while considering available information on parameter values and their levels of uncertainty. The uncertainty estimates that result are useful to decision makers determining the likelihood of different possible outcomes.

Although Bayesian calibration techniques have been developed relatively recently, they have been applied enough to demonstrate their utility and build a solid set of accepted practices. They have also been tested for use in building models, and have been found to be effective. The groundwork is now there for more widespread adoption, but challenges remain with regard to making Bayesian calibration more accessible. This paper aims to takes a step in that direction by providing some of the information needed to facilitate this adoption.

ACKNOWLEDGMENT

We would like to thank Yeonsook Heo for helping to start this research at Argonne and for much of the data used in this paper. We would also like to thank Dave Higdon and James Gattiker for their invaluable help in using and adapting the GPM/SA package for use with building energy models. This document has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. Work was supported by the DOE Building Technologies Office.

NOMENCLATURE

\( \theta \): parameters
\( x \): observed inputs
\( y \): model outputs
\( \eta(x, \theta) \): model emulator
\( \delta(x) \): model inadequacy
\( c(x, x') \): covariance matrix of \( x, x' \)
\( \sigma, \omega \): covariance hyperparameters

REFERENCES


© 2014 ASHRAE (www.ashrae.org). For personal use only. Reproduction, distribution, or transmission in either print or digital form is not permitted without ASHRAE’s prior written permission.