Calibrating Detailed Building Energy Simulation Programs with Measured Data—Part I: General Methodology (RP-1051)

T. Agami Reddy, PhD, PE Member ASHRAE
Itzhak Maor, PhD, PE Member ASHRAE
Chanin Panjapornpon, PhD

Received December 21, 2005; accepted July 31, 2006

Calibrated simulation is the process of using a building simulation program for an existing building and "tuning" or calibrating the various inputs to the program so that predictions match closely with observed energy use. Historically, the calibration process has been an art form that inevitably relies on user knowledge, past experience, statistical expertise, engineering judgment, and an abundance of trial and error. Unfortunately, despite widespread interest in the professional community, no consensus guidelines have been published on how to perform a calibration using detailed simulation programs. This research project was initiated with the intention to cull the best tools, techniques, approaches, and procedures from the existing body of research and develop a coherent and systematic calibration methodology that includes both parameter estimation and the determination of the uncertainty in the calibrated simulation. A general methodology of calibrating detailed simulation programs to performance data is proposed, which we deem to be methodical, rational, robust, and computationally efficient while being flexible enough to satisfy different users with different personal preferences and biases.

The methodology involves various concepts and approaches borrowed from allied scientific disciplines that are also reviewed in this paper. The methodology essentially consists of five parts: (1) identify a building energy program that has the ability to simulate the types of building elements and systems present and set up the simulation input file to be as realistic as possible; (2) depending on the building type, heuristically define a set of influential parameters and schedules that have simple and clear correspondence to specific and easy-to-identify inputs to the simulation program, along with their best-guess estimates and their range of variation; (3) perform a coarse grid search wherein the heuristically defined influential parameters are subject to a Monte Carlo simulation involving thousands of simulation trials from which a small set of promising parameter vector solutions can be identified by filtering, the strong and weak parameters can be identified, and narrower bounds of variability of the strong parameters can be defined; (4) perform a guided grid search to further refine the promising parameter vector solutions; and (5) use this small set of solutions (as opposed to a single calibrated solution) to make predictions about intended changes to the building and its systems, and determine the prediction uncertainty of the entire calibration process. A companion paper (Reddy et al. 2007) will present the results of applying this calibration methodology to two synthetic office buildings and one actual office building.

T. Agami Reddy is a professor in the Department of Civil, Architectural and Environmental Engineering, Drexel University, Philadelphia, PA. **Itzhak Maor** is a principal at PWI Energy Inc., Philadelphia. **Chanin Panjapornpon** is an assistant professor in the Department of Chemical Engineering, Kasetsart University, Bangkok, Thailand.

INTRODUCTION AND OBJECTIVE

Calibrated simulation as applied to building energy simulation computer programs is the process of "tuning" or calibrating the various inputs to the program so that observed energy use matches closely with that predicted by the simulation program. Though several practitioners perform calibrated simulations, most do so based on certain procedures they have developed over years of experience. There seems to be a lack of homogeneity among these procedures and, more importantly, a lack of proper appreciation of the procedures followed by calibration professionals as a whole. Thus, historically, calibration has been an art form that inevitably relies on user knowledge, past experience, statistical expertise, engineering judgment, and an abundance of trial and error. To date, no consensus guidelines have been published on how to assess the comparison of the results from a building energy simulation program against measured data from an actual building. ASHRAE initiated a research project (RP-1051) intended to cull the best tools, techniques, approaches, and procedures from the existing body of research and develop a coherent and systematic calibration methodology that includes both parameter estimation and determination of the uncertainty in the calibrated simulation (Reddy et al. 2006). The objectives of this research also involved studying the extent to which an accurate calibration fit to the billing data subsequently allows accurate prediction of energy use under operational and/or equipment changes to the existing building and developing a procedure following which the associated prediction uncertainty can be specified.

A previous paper (Reddy 2006) provided a literature review of calibrated simulation techniques, describing their uses, strengths, weaknesses, procedures, tools, and pertinent issues related to model fitting uncertainty. This paper will present a general methodology of calibrating detailed building energy simulation programs, along with guidelines, tips, and recommendations that will allow practitioners to use their preferred procedures (if they so deem) in a more consistent and scientific manner within a well-structured overall framework. Though the methodology is applicable to any building energy simulation program, the scope of RP-1051 was *restricted to the DOE-2 program*, which is a widely used public domain fixed schematic hourly simulation program (Winkelmann et al. 1993), and to the widely prevalent case where *year-long utility billing data* are the only performance data available for calibration. Moreover, it was presumed that the level and accuracy of knowledge about the building geometry, scheduling, and various system equipment would be consistent with a "detailed investment grade" audit, involving equipment nameplate information as well as some limited on-site measurements (clamp-on meters, etc.) performed during different times of the day (morning, afternoon, night) as well as over different days of the week in order to better understand variability in some of the simulation inputs. Finally, it was determined that the scope of the research was most pertinent to medium and large commercial buildings with relatively complex HVAC&R equipment that can be modeled by the simulation program to be calibrated. It is only in such buildings that the cost of performing calibrated simulations would be justified. Energy-saving measures in small buildings are relatively easy to identify, and neither the audit budget nor the monitoring and verification (M&V) budget may be large enough to allow performing calibration simulation.

The RP-1051 research was also meant to benefit software developers in that it would specify additional capabilities to existing building energy simulation programs, which would allow calibration to be performed by practitioners with relative ease and with higher consistency. A companion paper (Reddy et al. 2007) will present the results of applying the calibration methodology to both synthetic and actual case study buildings. The appendices in this paper include literature reviews of several concepts, issues, and procedures related to the building design process that were deemed pertinent enough to the calibration process that they would be of interest to calibration professionals.

Review of Allied Science and Engineering Literature on Calibration

Model calibration and validation have been addressed in several books and journal articles in allied areas of engineering and science, such as environmental, structural, hydrological, epidemiological, and structural engineering. The problem, which has been well known in these fields for several decades, is similar to the problem of calibrating detailed building energy simulation programs, which is aptly stated by Hornberger and Spear (1981): "most simulation models will be complex, with many parameters, state-variables and non-linear relations. Under the best circumstances, such models have many degrees of freedom and, with judicious fiddling, can be made to produce virtually any desired behavior, often with both plausible structure and parameter values." This process is also referred to in the scientific community as *GIGOing* (garbage in–garbage out), where a false sense of confidence can result since precise outputs are obtained by arbitrarily restricting the input space (Saltelli 2002). A brief review of some salient background notions is given below.

First, we distinguish between two types of models: (1) single models, which can be mechanistic differential equations or even empirical correlations, or (2) linked complex models, which require simulation programs with relatively long run times. The literature abounds with calibration and parameter estimation studies pertinent to the first type of model, while those dealing with the second type are relatively few. This research is intended to deal primarily with the latter that include calibration of simulation models. Given the limited monitored data available, one can at best identify only some of the numerous input parameters on which a simulation program is based. Hence, a first important step is to perform a sensitivity analysis in order to identify the influential system or model parameters (scientists often view this in terms of uncertainty, i.e., which are the main contributors to model output uncertainty).

There is a wide range of such analytical methods presented in a book edited by Saltelli et al. (2000) and a very large number of technical papers. In such methods, the analyst is often faced with the difficult task of selecting the one method most appropriate for a particular application. Broadly speaking, sensitivity analysis methods can be viewed as mathematical, statistical, or graphical. A workshop on sensitivity analysis was hosted in 2001, where selected experts were invited to write and present white papers reviewing the application of sensitivity and/or uncertainty analysis to complex engineering and/or environmental systems (Frey 2002; Frey and Patil 2002; Saltelli 2002).

A report by Iman and Helton (1985) compares different sensitivity analysis methods as applied to complex engineering systems and summarizes current knowledge in this area. Of all the techniques, three have been found to be promising: (1) *response surface* replacement of the computer model where fractional factorial design is used to generate the response surface—this method is optimal if the models are linear; (2) *differential analysis*, which is intended to provide information with respect to small perturbations about a point—however, this approach is not suited for complex models with large uncertainties; (3) *Latin hypercube sampling*, which was deemed very suitable overall to the current research. A short introduction to this method follows.

Experimental design methods, such as 2^k or 3^k factorial designs, have been in existence for several decades, well before the computer era. However, these methods have not been identified as promising since they only provide one-way sensitivity (i.e., the effect on the system response when only one parameter is varied at a time) rather than the multi-response sensitivity sought. When the number of input parameters is large, along with large uncertainty in the input parameters, and when the input parameters are interdependent and nonlinear, Monte Carlo (MC) methods (though computationally more demanding) are simpler to implement and require a much lower level of mathematics while providing adequate robustness.

In general, theoretical mathematicians deduce conclusions from postulates, while experimental mathematicians infer conclusions from observations (Hammersley and Handscomb 1964). MC methods, of which there are several types, as discussed below, comprise that branch of experimental mathematics that rely on experiments, using random numbers to infer the response of a system. Two types of problems can be handled by MC methods—probabilistic and deterministic—according to whether or not they are directly concerned with the behavior and outcome of random processes. For example, MC methods can be used to infer sampling distributions (e.g., mean, interquartile ranges). They can also be used as a numerical method of solving engineering problems using random sampling wherein certain determining factors are either obscure or subject to random processes. Essentially, MC methods are based on random sampling where numerous scenarios (on the order of hundreds to thousands) or different combinations of driving factors are simulated. The advantages of MC methods are (1) a low level of mathematics, (2) applicability to a large number of different types of problems, (3) that they can account for correlations between inputs, and (4) that can be applied to an unknown distribution of model variables. MC methods have been used extensively in risk assessment studies. Since MC methods are numerical methods, the relatively large computational effort may be viewed as a detriment by some; however, these methods are relatively efficient compared to other alternatives. The concept of "efficiency" has been used to compare different methods of implementing MC methods (Saltelli et al. 2000). Say two methods, methods 1 and 2, are to be compared. Method 1 calls for n_1 units of computing time (i.e., the number of times that the simulation is performed), while method 2 calls for n_2 times. If the resulting estimates of the response variable have variances σ_1^2 and σ_2^2, then the efficiency of method 2 with respect to method 1 is defined as

$$\frac{\varepsilon_1}{\varepsilon_2} = \frac{n_1 \cdot \sigma_1^2}{n_2 \cdot \sigma_2^2}, \tag{1}$$

where (n_1/n_2) is called the labor ratio and (σ_1^2/σ_2^2) is called the variance ratio.

Finally, MC methods have emerged as a basic and widely used tool to quantify uncertainties associated with model predictions and also for examining the relative importance of model parameters in affecting model performance (Spears et al. 1994). There are different types of MC methods depending on the sampling algorithm of generating the trials (Helton and Davis 2003):

- *Hit and Miss MC methods*, which were the historic manner of explaining MC methods. They involve using random sampling for estimating integrals (i.e., for computing areas under a curve and solving differential equations).
- *Crude MC method* using traditional random sampling, where each sample element is generated independently following a pre-specified distribution.
- *Stratified MC method* (also called *importance sampling*), where the population is divided into groups or strata according to some pre-specified criterion and sampling is performed so that each strata is guaranteed representation (unlike the crude MC method). This method is said to be an order of magnitude more efficient than the crude MC method.
- *Latin Hypercube MC (LHMC) method*, which uses stratified sampling without replacement and is easiest to implement, especially when the number of variables is large. It can be viewed as a compromise procedure combining many of the desirable features of random and stratified sampling. It produces more stable results than random sampling and does so more efficiently. It is easier to implement than stratified sampling for high-dimension problems since it is not necessary to determine strata and strata probabilities. Because of its efficient stratification process, LHMC is said to be one of the most promising methods for performing sensitivity

studies in long-running complex models, such as those used in building energy simulations (for example, Hofer [1999]).

LHMC sampling is conceptually easy to grasp. Say a sample of size n is to be generated from $\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_n]$. The range of each variable x_j is divided into n disjoint intervals of equal probability, and one value is selected randomly from each interval. The n values thus obtained for x_1 are paired at random without replacement with similarly obtained n values for x_2. These n2-pairs are then combined in a random manner without replacement with the n values of x_3 to form n3-triples. This process is continued until a sample of $n\boldsymbol{x}$-tuples is formed. This constitutes one LHMC sample or trial. How to modify this method to deal with correlated variables has also been proposed. A paper by Helton and Davis (2003) cites over 150 references in the area of sensitivity analysis, discusses the clear advantages of LHMC sampling for analysis of complex systems, and enumerates the reasons for the popularity of such methods.

Spears et al. (1994) deal with estimation of risk arising from the release of toxic chemicals from hazardous waste sites. Though the application area is very different from building energy modeling, the approaches and techniques for calibration are surprisingly relevant. An appropriate technique is regional sensitivity analysis (RSA), which involves (a) partitioning the parameter space of complex models with a large number of parameters into small, densely populated regions and relatively large, sparsely populated regions, wherein the former allow identification of the important controlling parameters, and (b) categorizing them in a statistical manner as acceptable or not.

Saltelli and Marivoet (1990) compare a number of different sensitivity analysis techniques in the case of nonlinear model responses. The test models are in the context of risk analysis for the disposal of radioactive waste where sensitivity plays a key role. Ten different statistical measures (Pearson correlation coefficient, partial correlation coefficient, standardized regression coefficient, Smirnov test statistic, Mann-Whitney test statistic, Spearmann rank correlation coefficient, partial rank correlation coefficient, standardized rank regression coefficient, Craner–von Mises test statistic, and two-sample t-test statistic) are applied to the output from the same MC simulations where random sampling is used for sample selection. Hypothesis testing is systematically applied to quantify the degree of confidence in the results given by the various sensitivity estimators. Although the relative efficiency is not studied, the estimators are ranked according to their robustness and stability for the case studied.

Eisenberg et al. (1996) developed a quantitative approach by using probability distributions rather than point estimates to characterize the microbial risk that an individual will become affected. A dynamic model was proposed along with a simulation methodology that explicitly considered the uncertainty and variability associated with the data. In this study, an MC approach, along with binary classification to assess the output of the simulations, was demonstrated by way of a case study example. The study showed that the MC simulation technique was able to identify conditions in which outbreaks were likely and identified parameters that contributed the most to the uncertainty associated with a risk prediction. A short literature review of relevant studies relating to the sensitivity of input parameters during the building design process is offered in Appendix A.

Proposed Methodology

The advocated methodology of calibrating a building simulation program against its utility billing data needs to be methodical, rational, robust, and computationally efficient. Further, it needs to be as inclusive as possible—i.e., it has to have the flexibility to cater to different users with different personal preferences and biases. For example, some analysts rely more on heuristics gained over several years, while others may wish to use a more mathematical approach. In

any case, it is clear from previous researchers (such as Carroll and Hitchcock [1993]) that calibrating a detailed energy simulation program involving numerous input parameters is a highly underdetermined problem (i.e., the presence of too many parameters is likely to result in any solution being non-unique).Thus, a satisfactory overall calibration to the utility billing data will not guarantee accurate identification of the individual parameters in the simulation program. Our calibration methodology explicitly recognizes the fact that it is unlikely that any one calibrated solution (defined by one vector of input parameter values) can be deemed the "best" solution. Hence, our approach advocates that it is much more robust to identify a small set of most plausible solutions instead. This is consistent with Saltelli (2002), who, in the framework of calibration in traditional scientific disciplines, urges the analyst to refrain from searching for an optimal solution and instead be satisfied with several plausible ones. Further, Oliva (2003) points out that a good fit to the data, though a first requirement, is not a sufficient condition when calibrating a model and that full consideration of the estimated parameters as they shape the model structure should be explicitly considered. Hence, the general calibration methodology we advocate involves the following five major steps (see Figure 1):

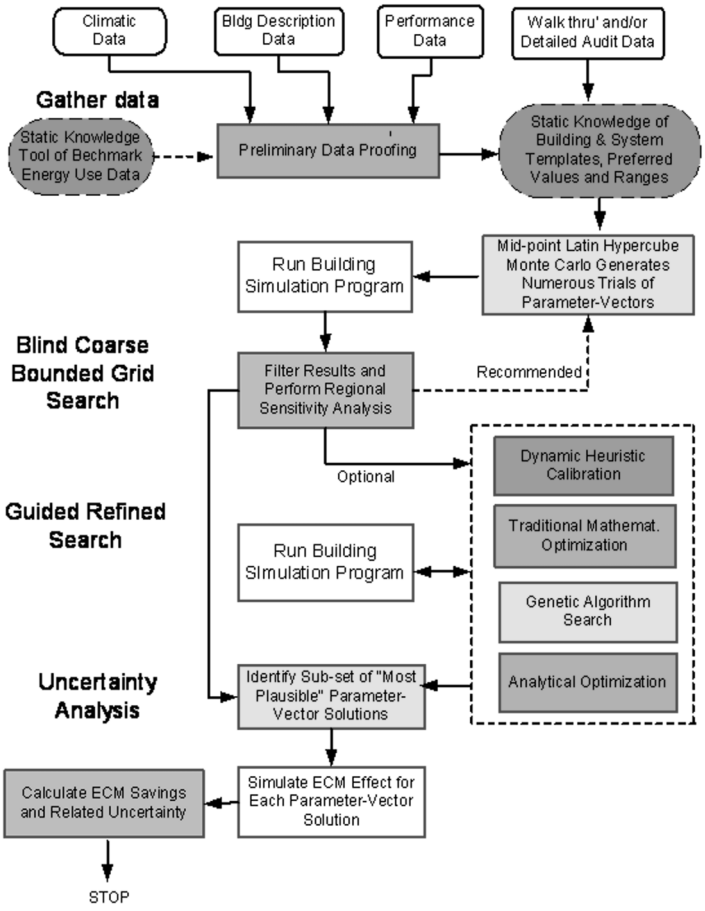

Figure 1. Structure of calibration methodology proposed.

1. An important first step is to prepare a preliminary simulation input file of the building that is as realistic and error-free as possible. This would entail making sure that the simulation program has the capability of modeling the type of building and systems present and that the inputs have been entered correctly.
2. Next, reduce the dimensionality of the parameter space by resorting to walk-through audits and heuristics. For a given building type, identify/define a set of influential parameters and building operating schedules along with their best-guess estimates (or preferred values) and their range of variation characterized by either the minimum–maximum range or the upper and lower 95th probability threshold values. The set of influential parameters to be selected should be such that they correspond to specific and easy-to-identify inputs to the simulation program.
3. Next, perform a "bounded" coarse grid calibration (or unstructured or blind search) using an MC simulation involving numerous trials or realizations with different combinations of input parameter values. This would result in a preliminary filtering or the identification of a small set of the most promising solutions of input parameter vectors and also provide a means of identifying the more sensitive or strong parameters.
4. Subsequently, perform a guided search calibration to further refine or improve on the calibrated solutions identified by the coarse grid search. In the companion paper (Reddy et al. 2007), we will evaluate whether the guided search phase is necessary in the first place or whether a blind bounded search is adequate to provide plausible solutions of acceptable accuracy.
5. Finally, rather than using only one plausible calibrated solution to make predictions about the effect of intended energy conservation measures (ECMs), use a small number of the most plausible solutions. Not only is one likely to obtain a more robust prediction of the energy and demand reductions, but this would allow determining their associated prediction uncertainty as well.

Elements of the Calibration Methodology

Figure 1 is a flowchart depicting the various elements that comprise the various steps stated above and their interactions in the framework of the proposed calibration methodology. The major elements and their subelements are expanded below.

1. **Gathering baseline information and proofing of performance data.** This is an important preparatory stage and will involve gathering and checking the quality of both energy use data to be used for calibration as well as the building and system specifications gathered from as-built drawings and detailed walk-through site visits. This element would involve:

   (a) *Checking performance data indices* (such as annual energy use per square foot) against benchmark energy use data (captured in a knowledge base) to detect gross errors. The Commercial Building Energy Survey (CBES) database developed by the US Energy Information Administration (USEIA 1995) has been extensively used by several researchers to develop energy and demand benchmarks in the forms of Energy Use Intensities (EUIs) for different types of commercial buildings (for example, Burns [1990]; Eto [1990]; Sharp [1994]). A subsequent study by Huang and Franconi (1995) describes how the CBES data was used to develop generic prototype building data for 12 major building types and two different vintages in five different US locations that were then simulated using DOE-2 in order to generate knowledge relating to end-use energy densities for these buildings as well as quantify the relative contribution of building shell, secondary system, and plant. The effect of different types of secondary systems, such as constant-air-volume (CAV) and variable-air-volume (VAV) systems, was also considered.

(b) *Evaluating data quality* by means of visual or statistical screening methods (for example, Sonderegger et al. [2001] and ASHRAE Guideline 14-2002 [ASHRAE 2002]).

(c) *Creating audit data forms* for mandatory (or basic) and comprehensive information (depending on the level of data availability) about building and systems required to launch the DOE-2 simulation program. Several sources can be consulted for this purpose, for example, RESEM (Carroll and Hitchcock 1993), TRACE building templates (Trane 1992), the more than a dozen previous studies described by Huang and Franconi (1995), as well as the minimum requirements of the prescriptive path of ANSI/ASHRAE/IESNA Standard 90.1 (ASHRAE 2001).

2. **Define a heuristic static knowledge base of templates of the influential parameter inputs to the detailed simulation model, which will depend on the specific building and system type.** The list, though tailored to the building type and HVAC system used, has to be generic and may contain more parameters than those that will turn out to be strong for any specific circumstance. Carroll and Hitchcock (1993) used 23 influential input parameters, though they point out that reducing this number would be advisable. Note that the parameters can be either discrete or continuous. This element would involve:

   (a) *Developing DOE-2 consistent simulation input templates* generic to a building type and HVAC system using audit information.

   (b) *Isolating/separating influential and non-influential variables*, both discrete and continuous, depending on the particular circumstance, for electricity use and demand, and gas use. One can refer to previous papers and reports (see Reddy [2006]) combined with heuristic knowledge and experience. Since different types of systems may have different strong and weak variables, static heuristic templates have to be specific to the specific case considered.

   (c) *Defining preferred or base-case input values for influential variables* involves assigning lower and upper ranges (or minimum and maximum values) to numerical values for continuous variables and specific values for discrete variables (for example, diurnal lighting schedule code).

3. **Use a blind coarse grid approach to sample the search space in a random manner.** We adopt a blind bounded search method by first discretizing the range of variability of the continuous parameters so as to reduce the number of combinations (or the size) of the search space and then using a random sampling method. More specifically, this module is meant to identify promising regions of local optima by using the LHMC method. Specifically, the following steps are required:

   (a) *Define "best guess" default values for non-influential variables* that do not change during the entire calibration process using knowledge specific to the building.

   (b) *Discretize the probability distribution* or variability of the continuous variables into strata or levels. Though using a two-level approach reduces the number of trials by several orders of magnitude, we have chosen to use three levels since such an experiment can account for the more realistic situation of nonlinear behavior between energy use and the various parameters (a two-level experiment would force this behavior to be linear).

   (c) *Determine probability-weighted, mid-point values for each strata*. Even for the simpler case, the heuristic templates may consist of 20 or more input parameters. Generating suitable combinations using traditional factorial design or response surface designs are grossly inadequate, and so the LHMC sampling of the influential variables is most

appropriate (see Appendix A for necessary background). We suggest assuming triangular probability distributions for capturing the variability in the influential parameters (see Figure 2). The minimum, base-case, and maximum values of the various parameters can be used to determine three discrete sub-ranges, called low, mid, and high values from which probability-weighted mid-point values are calculated (see report by Reddy et al. [2006] for detailed algorithms). It must also be stated that adopting a triangular distribution is sound if we have a certain amount of confidence in our best guess or base values. If we do not (i.e., if we only have confidence in the range of variation of the parameter in question), then using an uniform distribution, which is easier to implement, is recommended.

(d) *Generate the necessary number of combinations or trials* of the numerous influential input parameters. Because of the relatively large number of initial parameters to be considered, there are a large number of combinations—called the "curse of dimensionality." For example, the number of combinations M of 20 parameters p with each at three different levels is $M = 3^{20} \approx 3.5 \times 10^{9}$. Reducing the number p by five would result in $M = 3^{15} \approx 1.4 \times 10^{7}$, which is more than a two orders of magnitude reduction. Further, reducing p to 10 would result in $M = 3^{10} \approx 6 \times 10^{4}$, again more than a two orders of magnitude reduction. Hence, reducing the dimensionality of the search is of critical importance. Note that with $p = 20$, performing even 3500 batch simulation runs would result in sampling the search space once in a million possibilities, which is a very sparse sampling process indeed.

(e) *Run DOE-2 simulation program* in batch mode for as many trials as generated above.

(f) *Filter the large number of batch LHMC trials*, i.e., identify promising subsets of parameter vector combinations based on goodness-of-fit criteria between simulated and actual monthly utility bill data over the year (described in the next section).

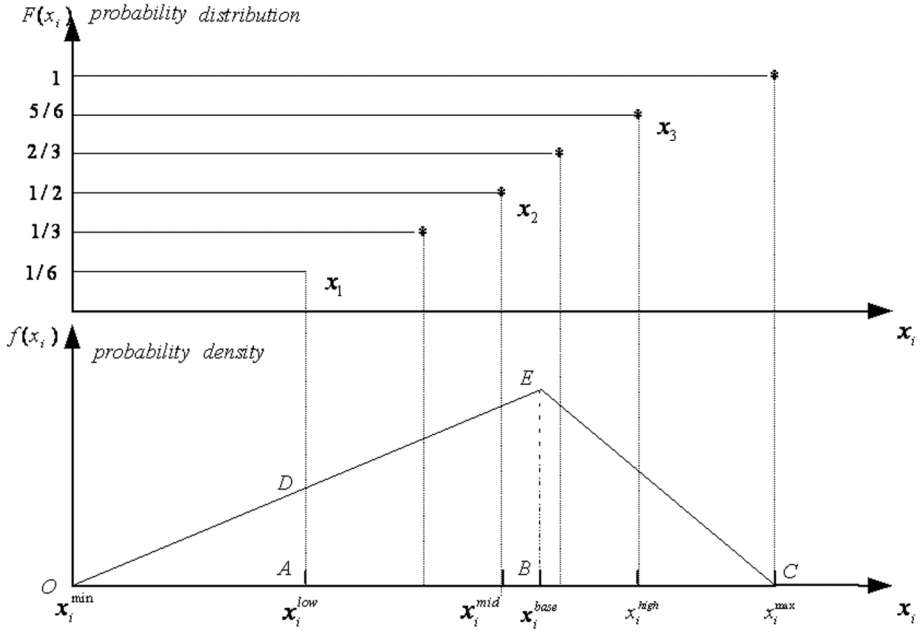

Figure 2. Sketch illustrating how the continuous parameters are discretized assuming a triangular distribution.

(g) *Perform sensitivity analysis* using the above subset of promising trials to identify the subset of strong parameters from the list of influential input parameters. The next section provides specific details about such goodness-of-fit (GOF) statistical criteria.

(h) *Perform replicate tests* to obtain statistically robust results, if necessary, by repeating steps (c) and (d) with another set of random runs to verify whether the same (or at least some of the) strong parameters and their numerical ranges are obtained. One efficient option is to "freeze" the weak variables so as to reduce the dimensionality of the search space and let only the remaining variables assume random values in accord with the LHMC algorithm.

4. **Adopt a refined, guided search, which will start with each of the top "promising" solutions for the input parameters from the blind coarse grid search and systematically fine-tune or refine the numerical values of the strong parameters so as to improve the prediction accuracy of the simulation outputs to the utility bill data as closely as possible**. What differentiates this step is that it starts with the estimates of the strong input parameters identified from the blind coarse grid search and incrementally improves the calibration to reach closer match with the utility bills. The guided search process could be one of several approaches, such as the manual, iterative, and pragmatic intervention approach described in Reddy (2006), or mathematical, depending on the personal preference or capabilities of the analyst. While the former will involve using rules derived from the past experience and expertise of the user, the mathematical approach could use any of the traditional optimization methods such as gradient-based or non-gradient-based methods (for example, Cumali [1988], Reinhard and Duffy [2001], Christensen et al. [2003]). Nontraditional optimization methods, such as simulated annealing or genetic algorithms (for example, Wetter [2001], Caldas and Norford [2003]), could even be adopted. Since the coarse grid search identifies a number of promising solutions or feasible input parameter vectors, we can use this information advantageously by assuming these to be the initial population for initiating the genetic algorithm search. This would improve the efficiency of the entire search process and would appeal to analysts who would like to use additional automated calibration capability instead of heuristic calibration. However, this approach is computationally demanding. A third option involves adopting purely analytical methods to perform this final calibration process (Sun and Reddy 2006).
5. **Identify plausible sets of solutions or plausible input parameter vectors.** As discussed previously, because calibration is a highly underdetermined problem, more than one solution can satisfy the objective function, i.e., provide good fits to utility bill (or monitored) data. Hence, we propose that a set of plausible strong input parameter solutions rather than a single solution be deemed the "best" solution. This is consistent with the suggestions by Saltelli (2002) and Oliva (2003) stated earlier. The plausible parameter vector sets can be directly identified either from the coarse grid search or from the guided refined search. The case studies presented in the companion paper (Reddy et al. 2007) will help us investigate this aspect.
6. **Compute uncertainty in calibrated model prediction.** The final step involves proposing a statistically sound approach along with recommendations and suggestions. This is an important step in our methodology given that this issue has been entirely overlooked by practitioners of calibrated solutions. The conventional wisdom is that once a simulation model is calibrated with actual utility bills, the effect of different intended ECMs can be predicted with some degree of confidence by making changes to one or more of the model input parameters that characterize the ECM. Such thinking is clearly erroneous. The utility billing data is the aggregate of several end-uses within the building, each of which is affected by one or more specific and interacting parameters. While performing calibration, the many degrees of

freedom may produce good calibration overall even though the individual parameters may be incorrectly identified. Subsequently, altering one or more of these incorrectly identified parameters to mimic the intended ECM is very likely to yield biased predictions. Our approach to overcoming this problem is discussed in the following section.

Statistical Criteria of Determining Goodness-of-Fit

Usually utility bills include electricity use in kWh, demand in kW,[1] and gas use,[2] and calibration involves reconciling all three quantities. Traditionally, the two most often used statistical indices for GOF are the normalized mean bias error (NMBE) and the coefficient of variation (CV), which are akin to the mean and the standard deviation of the residuals (ASHRAE 2002). During calibration, one needs an aggregated index that will be indicative of how well the calibrated model fits all these three types of energy use and allow easier ranking of the various simulation results during the calibration process. There are different ways of developing such an aggregated index, but the most intuitive is to weight them individually depending on how they contribute to the annual energy cost. Such a financially weighted GOF index can be deduced as follows (Kreider and Haberl 1994; Sonderegger et al. 2001):

$$\mathrm{GOF}_{\mathrm{CV}} = \left[\frac{(w_{\mathrm{kWh}}^2 \mathrm{CV}_{\mathrm{kWh}}^2 + w_{\mathrm{kW}}^2 \mathrm{CV}_{\mathrm{kW}}^2 + w_{\mathrm{th}}^2 \mathrm{CV}_{\mathrm{therms}}^2)}{(w_{\mathrm{kWh}}^2 + w_{\mathrm{kW}}^2 + w_{\mathrm{th}}^2)}\right]^{1/2} \tag{2}$$

and

$$\mathrm{GOF}_{\mathrm{NMBE}} = \left[\frac{(w_{\mathrm{kWh}}^2 \mathrm{NMBE}_{\mathrm{kWh}}^2 + w_{\mathrm{kW}}^2 \mathrm{NMBE}_{\mathrm{kW}}^2 + w_{\mathrm{th}}^2 \mathrm{NMBE}_{\mathrm{therms}}^2)}{(w_{\mathrm{kWh}}^2 + w_{\mathrm{kW}}^2 + w_{\mathrm{th}}^2)}\right]^{1/2} \tag{3}$$

where w_{kWh} is a weighted ratio that can be chosen as the ratio of the annual cost of electricity use in kWh divided by the total annual utility cost.

If the simulation results are plotted on a x-y plot with the above two indices (which are normalized indices) on the two axes, we would like to identify the simulation points whose distance is closest to the origin as "better" calibrated than those farther away from the origin. Hence, we introduce another set of weights, w_{CV} and w_{NMBE}, which impact our consolidated index as follows:

$$\mathrm{GOF\_Total} = \left[\frac{(w_{\mathrm{CV}}^2 \mathrm{GOF}_{\mathrm{CV}}^2 + w_{\mathrm{NMBE}}^2 \mathrm{GOF}_{\mathrm{NMBE}}^2)}{(w_{\mathrm{CV}}^2 + w_{\mathrm{NMBE}}^2)}\right]^{1/2} \tag{4}$$

where $(w_{\mathrm{CV}} + w_{\mathrm{NMBE}}) = 1$. Practically, building energy managers would prefer the calibration to capture the mean more accurately than the month-by-month variation (described by the CV); this feeling is implicitly captured by ASHRAE Guideline 14-2002 (ASHRAE 2002), which assigns a 1:3 weight for (w_{CV}: w_{NMBE}).

[1] The actual electric demand is usually charged based on either 15-minute or 30-minute usage. However, most of the building energy simulation programs implicitly assume hourly simulation time steps, so reconciliation of the measured and simulated demands contains this inherent mismatch, which introduces a certain amount of error during calibration.

[2] Gas use is charged in terms of volume consumed. However, we shall use units of heat to express gas use, i.e., therms or kBtu, so as to be consistent with electricity use.

Sensitivity Analysis to Identify Strong and Weak Parameters

The previous section dealt with the statistical criteria of determining how well the calibrated simulation model matched the actual building energy performance. MC filtering is the process of rejecting sets of model simulations that fail to meet some prescribed criteria of model performance; such criteria were previously suggested. Sensitivity analysis goes a step further. As discussed in the literature review (Appendix A), the objective of a sensitivity analysis in our context is to identify the parameters that are strong (i.e., those that influence the statistical GOF criteria denoted by Equations 2–4) and those that are weak among the influential parameter sets. Once these are determined, the values of the "weak" parameters can be fixed at their nominal values, thus reducing the number of parameters to be calibrated. This would reduce the parameter space enormously and somewhat alleviate the "curse of dimensionality."

Since running detailed building energy simulation programs is computationally intensive and they have long runtimes, one cannot afford to perform separate simulation runs for identifying promising parameter vector combinations and for sensitivity analysis. Therefore, we propose the following procedure in order to satisfy both these objectives simultaneously.

Once an LHMC batch run consisting of, say, 1000 trials is completed, the GOF_CV and GOF_NMBE indices are computed for each trial. From here we can filter or weed out those parameter vectors that result in high GOF numbers, (i.e., those whose predictions fit the utility bills poorly). One suggestion is to retain all those parameter vectors that resulted in GOF_NMBE < 10% and GOF_CV < 20% (ASHRAE Guideline 14-2002 suggests 5% and 15%, respectively, when utility bills are being reconciled). The information contained in these promising parameter vectors can be used to identify the weak parameters, which can then be removed from further consideration in the calibration process.

Assume that 30 candidate vectors were identified from the coarse grid search. Since each parameter has been discretized into three states, we would expect that if the individual parameters were weak they would be randomly distributed in these 30 candidate parameters. Thus, the extent to which the number of occurrences of an individual parameter differs from ten within each discrete state would indicate whether it is strong or weak. This is a type of sensitivity test where the weak and strong parameters are identified using nonrandom pattern tests (Saltelli et al. 2000). The Chi-square (χ^2) test is advocated as a test for statistical independence for each and every parameter. This is a well-known test for comparing distributions (see any statistics textbook) whereby the χ^2 statistic is first computed as follows:

$$\chi^2 = \sum_{s=1}^{3} \frac{(p_{\text{obs},s} - p_{\text{exp}})^2}{p_{\text{exp}}} \tag{5}$$

where p_{obs} is the observed number of occurrences and p_{exp} is the expected number (in our example above, this will be ten) and the subscript s refers to the number of states or the degrees of freedom (three states are assumed in this research).

If the observed number is close to the expected number, the χ^2 value will be small, indicating that the observed distribution fits the theoretical distribution closely. This would imply that the particular parameter is weak since the corresponding distribution can be viewed as being random. Note that this test requires that the degrees of freedom (d.f.) be selected as (number of states minus one), i.e., in our case, d.f. = 2. The critical values for the χ^2 distribution for different significance levels α are given in Table 1.

Table 1. Critical Thresholds for the Chi-Square Statistic

d.f.	$\alpha = 0.001$	$\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.9$
2	13.815	10.597	9.210	5.991	3.219	2.408	1.386	0.211

Selecting these critical values as dictating the strength of any parameter is to some extent arbitrary and needs to be tested with synthetic data (which will be addressed in the companion paper, Reddy et al. [2007]). If the χ^2 statistic for a particular parameter is greater than 9.21, one could safely assume it to be very strong since the associated statistical probability is greater than 99%. On the other hand, a parameter having a value of 1.386 ($\alpha = 0.5$) could be considered weak and those in between the two values as uncertain in influence.

As a final note, we advocate the use of Table 1 to determine whether the LHMC trials generated by the computer are really random or not. Since computers generate quasi-random numbers, we have found that it is essential to explicitly program the LHMC trial generation algorithm to verify randomness so that the trials generated are really random in each of the parameters being calibrated. Otherwise, we could reach erroneous conclusions regarding the strength or weakness of individual parameters, especially in cases where one is running a batch with a relatively small number of trials.

Uncertainty Issues in Calibrated Model Prediction

Any comparison between measured and predicted performance should be seen not in terms of whether the program agrees with the measurements but *whether the program is good enough for its intended purpose.* The issue is, then, how does one determine a "good enough" threshold? The statistical basis of the manner in which ASHRAE Guideline 14-2002 addresses uncertainty issues is based on a methodology proposed by Reddy and Claridge (2000). This methodology applies to regression models identified from baseline monitored data and, hence, relate to black-box and grey-box approaches. They cannot be applied as such to the calibrated simulation model approach, which was one of the reasons why this research was undertaken in the first place.

Most fundamental of all, there is no broad consensus as to how a guideline document such as ASHRAE Guideline 14-2002 should stipulate the needed accuracy limits for the calibration to be deemed satisfactory. These limits may not be absolute limits but would depend on the intended application in which the calibrated simulation is likely to be used subsequently. If that is so, guideline documents should specify the methodology or process by which such accuracy limits can be determined for the specific circumstance when the calibrated simulation approach is adopted. Though the ASHRAE Guideline 14-2002 tolerance or uncertainty levels stated above may provide some sort of initial guidance, a more subtle approach to determining uncertainty limits for calibrating simulation is warranted because of an important but implied assumption. It is assumed that if a model is calibrated to these tolerances, it is sufficiently close to physical reality in that the individual input values are realistic (i.e., the "knobs" are tuned properly). Then the accuracy in the evaluation of the effect of a certain ECM can be attributed simply to the internal model accuracy. *Just because the net effect of all the knobs yields a simulation output close to the one desired (i.e., measured), it is no guarantee that all the individual knobs are tuned correctly*. Thus, in order for us to be confident in the energy savings of a potential ECM, we should know (1) which are the important input knobs that impact energy savings due that particular ECM and (2) be extra careful that this set of knobs is tuned correctly even though the remaining non-influential ones may not be. This issue, we feel, has been overlooked entirely in ASHRAE Guideline 14-2002 by not explicitly separating external or internal error types (Reddy 2006). We have been able to locate one reference (Kaplan et al. 1990) that looked at this issue and concluded that "investigation to date on the calibration tolerance required to fulfill the research objectives indicates that this is probably unanswerable."

An important fact highlighted above is that the calibration process does not guarantee that each of the numerous parameters to the simulation model are "tuned" correctly. Because the problem is underdetermined, we may have obtained one particular parameter set combination (or one plau-

sible solution) that yields a "local" minimum within the search space. As pointed out by Kaplan et al. (1990), one can never hope to identify these parameters correctly (partly because we do not know what is "right"). In such a case, limiting ourselves to one plausible solution will be misleading since one would rightly question the prediction accuracy of the change in energy use predicted by the calibrated model. In other words, a satisfactory overall calibration to the utility bill will not guarantee accurate fit at the end-use level and may yield unsatisfactory prediction accuracy of the effect of the intended ECM. That is why using a small number of top feasible solutions is bound to be more robust than using only one feasible calibrated solution to make predictions about the effect of intended ECMs. What is deemed to be a "small number" is an issue that is addressed in the companion paper dealing with the case studies (Reddy et al. 2007).

Finally, there is the issue of uncertainty in the building energy use performance data against which the calibration is being performed. Often, it is tacitly assumed that such data are deterministic, involving no uncertainty at all. This is certainly not the case. For example, utility bills are often not read on the exact date and time each month. Reading the utility meter in the morning versus in the evening may result in 1%–2% differences in monthly values even if everything else is the same. If the instrument accuracy is folded in, the uncertainty levels could be 3% or higher. Such errors are even higher for energy use data collected from field monitoring equipment. Such issues are related to the robustness of the calibration methodology and have been addressed in the framework of the case studies in the companion paper (Reddy et al. 2007).

SUMMARY AND FUTURE WORK

In this paper, we have proposed a general methodology of calibrating detailed simulation programs to performance data, which we deem to be methodical, rational, robust, and computationally efficient while being flexible enough to satisfy different users with different personal preferences and biases. We started by emphasizing explicitly the fact that calibration is an underdetermined or overparametrized problem whose order has to be first reduced by using heuristic insight as well as acquiring detailed information from investment-grade audits and at least spot measurements. Even then, it is unlikely that any one calibrated solution (defined by one vector of input parameter values) can be deemed to be the "best" solution. Hence, our approach asserts that it is futile to search for an optimal solution and that it is much more robust to identify a small set of "most plausible solutions" instead. This research has also proposed and developed several important implementation issues that are novel to the area of calibrating detailed building energy simulation programs:

- the definition of a set of influential input parameters and schedules along with their best-guess estimates, their range of variation depending on the building type, and the underlying probability distribution (either triangular or uniform), which are then discretized into three ranges of equal probability;
- performing a coarse grid search wherein the heuristically defined influential parameters are subject to a LHMC simulation involving thousands of trials and filtering them so as to identify the strong and weak parameters, defining narrower bounds of variability of the strong parameters, and identifying a small set of feasible calibrated solutions of the input parameters that fit the data as closely as possible;
- performing a guided grid search to further refine the promising solutions, which can be done in one of several ways depending on the personal preference of the analyst; and
- using a small set of the top (or most promising) calibrated solutions to make more robust predictions of energy savings from intended ECMs and determining the associated prediction uncertainty.

A companion paper (Reddy et al. 2007) presents the results of applying this calibration methodology to synthetic and actual office buildings. The application of this methodology to the case study buildings supports the original intention of being able to automate the entire process and also apply it to cases where additional monitored performance data (short-term monitoring or interval data) are available. Further, several aspects of this research can be immediately adopted by those performing calibrated simulations, while commercial adoption would require some additional research into certain aspects identified in the final research report of RP-1051 (Reddy et al. 2006).

ACKNOWLEDGMENTS

This research was sponsored by ASHRAE TC 4.7, Energy Calculations. We would like to acknowledge the Project Monitoring Subcommittee members of this project, namely, Robert Sonderegger (chair), Bass Abushakra, Marlin Addison, Chip Barnaby, Ron Judkoff, and Vern Smith, for their timely, insightful, and constructive criticism and support during the course of this research. Several other colleagues provided advice, support, and help at different times related to different aspects of this research: David Binder, Drury Crawley, William Carroll, Charles Haas, Jeff Haberl, Srinivas Katipamula, Rob Ramirez, Jiang Sun, Peter Vance, James Waltz, and Jack Wolpert.

REFERENCES

ASHRAE. 2001. *ANSI/ASHRAE/IESNA Standard 90.1-2001, Energy Standard for Buildings Except Low-Rise Residential Buildings*. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE. 2002. *ASHRAE Guideline 14-2002, Measurement of Energy and Demand Savings*. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Box, G.E.P., W.G. Hunter, and J.S. Hunter. 1978. *Statistics for Experimenters—An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley & Sons.

Burns, E.M. 1990. Annual and monthly peak demand for electricity in commercial buildings. *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings* 3:47–55.

Caldas, L.G., and L.K. Norford. 2003. Genetic algorithms for optimization of building envelopes and the design and control of HVAC systems. *ASME Journal of Solar Energy Eng.* 125:343–51.

Carroll, W.L., and R.J. Hitchcock. 1993. Tuning simulated building description to match actual utility data: Methods and implementation. *ASHRAE Transactions* 99(2):928–34.

Chou, S.K., and W.L. Chang. 1993. Development of an energy-estimating equation for large commercial buildings. *International Journal of Energy Research* 17:759–73.

Chou, S.K., W.L. Chang, and Y.W. Wong. 1993. Effects of multi-parameter changes on energy use in large buildings. *International Journal of Energy Research* 17:885–903.

Christensen, C., B. Stoltenberg, and G. Barker. 2003. An optimization methodology for zero net energy buildings. *ASME Solar Energy Conference Proceedings, Hawaii, June.*

Corson, G.C. 1992. Input-output sensitivity of building energy simulations. *ASHRAE Transactions* 98(1):618–26.

Cumali, Z. 1988. Global optimization of HVAC system operations in real time. *ASHRAE Transactions* 94(1):1729–43.

De Wit, S., and G. Augenbroe. 2002. Analysis of uncertainty in building design evaluations and its implications. *Energy and Buildings* 34:951–58.

Eisenberg, J.N., E.Y.W. Seto, A.W. Olivieri, and R.C. Spear. 1996. Quantifying water pathogen risk in an epidemiological framework. *Risk Analysis* 16(4):549–63.

Eto, J.H. 1990. An investigation of the use of prototypes for commercial sector EUI analysis. *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings* 10:29–37.

Frey, H.C. 2002. Introduction to special section on sensitivity analysis and summary of NCSU/USDA workshop on sensitivity analysis. *Risk Analysis* 22(3):539–45.

Frey, H.C., and S.R. Patil. 2002. Identification and review of sensitivity analysis methods. *Risk Analysis* 22(3):553–78.

Hammersley, J.M., and D.C. Handscomb. 1964. *Monte Carlo Methods*. London: Methuen and Co.

Helton, J.C., and F.J. Davis. 2003. Latin hypercube sampling and the propagation of uncertainty of complex systems. *Reliability Engineering and System Safety* 81:23–69.

Hofer, E. 1999. Sensitivity analysis in the context of uncertainty analysis for computationally intensive models. *Computer Physics Communication* 117:21–34.

Hornberger, G.M., and R.C. Spear. 1981. An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management* 12:7–18.

Hou, D., J.W. Jones, B.D. Hunn, and J.A. Banks. 1996. Development of HVAC system performance criteria using factorial design and DOE-2 simulations. *Proceedings of the 10th Symposium on Improving Building Systems in Hot and Humid Climates, May 13–14, Fort Worth, TX*, pp. 184–92.

Huang, Y.J., J.W. Hanford, and M. Piraino. 1993. The impact of variations in building parameters and operating conditions on commercial building energy use and load shapes. *IBPSA Conference, Adelaide, Australia, Aug. 16–18*, pp. 215–22.

Huang, Y.J., and E. Franconi. 1995. Commercial building heating and cooling loads component analysis, LBNL-37208. Lawrence Berkeley National Laboratory, Berkeley, CA.

Hui, S.C.M. 1997. A randomized approach to multiple regression analysis of building energy simulation. *IBPSA Conference, Prague, Czech Republic, Sept. 13–1*5 1:103–10.

Iman, R.L., and J.C. Helton. 1985. A comparison of uncertainty and sensitivity analysis techniques for computer models. Report NUREG/CR-3904, SAND 84-1461, Sandia National Laboratories, Alberquerque, NM.

Judkoff, R., D. Wortman, B. O'Doherty, and J. Burch. 1983. A methodology for validating building energy analysis simulations, SERI/TR-254-1508. Solar Energy Research Institute (now National Renewable Energy Laboratory), Golden, CO.

Kaplan, M.B., J. McFerran, J. Jansen, and R. Pratt. 1990. Reconciliation of a DOE2.1C model with monitored end-use data for a small office building. *ASHRAE Transactions* 96(1):981–93.

Kreider, J.F., and J. Haberl. 1994. Predicting hourly building energy usage: The great predictor shootout—Overview and discussion of results. *ASHRAE Transactions* 100(2):1104–18.

Lam, J.C., and S.C.M. Hui. 1996. Sensitivity analysis of energy performance of office buildings. *Building and Environment* 31(1):27–39.

Lomas, K.J., and H. Eppel. 1992. Sensitivity analysis techniques for building thermal simulation programs. *Energy and Buildings* 19:21–44.

Macdonald, I., and P. Strachan. 2001. Practical application of uncertainty analysis. *Energy and Buildings* 33:219–27.

Mottillo, M. 2001. Sensitivity analysis of energy simulation by building type. *ASHRAE Transactions* 107(2):722–32.

Oliva, R. 2003. Model calibration as a testing strategy for system dynamics models. *European Journal of Operational Research* 151:552–68.

Reddy, T.A. 2006. Literature review on calibration of building energy simulation programs: Uses, problems, procedures, uncertainty, and tools. *ASHRAE Transactions* 12(1):177–96.

Reddy, T.A., and D.E. Claridge. 2000. Uncertainty of "measured" energy savings from statistical baseline models. *HVAC&R Research* 6(1):3–20.

Reddy, T.A., I. Maor, C. Panjapornpon, and J. Sun. 2006. Procedures for reconciling computer-calculated results with measured energy data, ASHRAE RP-1051, final report, January. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

Reddy, T.A., I. Maor, and C. Panjapornpon. 2007. Calibrating detailed building energy simulation programs with measured data—Part II: Application to three case study office buildings. *HVAC&R Research* 13(2).

Reinhard, M., and J. Duffy. 2001. Optimization methods for alternative energy system design. *Proceedings of the ASES Annual Conference, American Solar Energy Society, Boulder, CO*.

Saltelli, A. 2002. Sensitivity analysis for importance assessment. *Risk Analysis* 22(2):579–90.

Saltelli, A., and J. Marivoet. 1990. Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques. *Reliability Engineering and System Safety* 28:229–53.
Saltelli, A., K. Chan, and E.M. Scott (eds.). 2000. *Sensitivity Analysis*. Chichester: John Wiley and Sons.
Saporito, A., A.R. Day, T.G. Karayiannis, and F. Parand. 2001. Multi-parameter building thermal analysis using the lattice method for global optimization. *Energy and Buildings* 33:267–74.
Sharp, T.R. 1994. Energy benchmarking in commercial office buildings. *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings*, paper 4, pp. 321–29.
Sonderegger, R., J. Avina, J. Kennedy, and P. Bailey. 2001. Deriving loadshapes from utility bills through scaled simulation. Seminar presentation, ASHRAE Annual Meeting, Kansas City, MO.
Spears, R., T. Grieb, and N. Shiang. 1994. Parameter uncertainty and interaction in complex environmental models. *Water Resources Research* 30(11):3159–69.
Sullivan, R., S. Nozari, R. Johnson, and S. Selkowitz. 1985. Commercial building energy performance analysis using multiple regression. *ASHRAE Trans*actions 91(2):337–53.
Sun, J., and T.A. Reddy. 2006. Calibration of building energy simulation programs using the analytic optimization approach. *HVAC&R Research* 12(1):177–96.
Trane. 1992. TRACE 600 Program. Trane Company, LaCrosse, WI.
USEIA. 1995. Commercial buildings energy consumption and expenditures 1992. US Energy Information Administration, DOE/EIA-0318(92). US Energy Information Agency, Washington, DC.
Wetter, M. 2001. GenOpt: Generic Optimization Program. *Seventh International IBPSA Conference, Rio de Janeiro, Brazil*, pp. 601–608.
Winkelmann, F.C., B.E. Birdsall, W.F. Buhl, K.L. Ellington, A.E. Erdem, J.J. Hirsch, and S. Gates. 1993. DOE-2 Supplement, Version 2.1E, Lawrence Berkeley National Laboratory, Berkeley, CA, November.

APPENDIX A
PROCEDURES TO IMPROVE BUILDING DESIGN PROCESS

Several of the concepts, issues, and proposed procedures pertinent to the building design process are also applicable to the calibration process. In this appendix, published literature on building design relevant to identifying the sensitivity of input parameters, as well as methods to reduce the number of design simulation runs, are briefly reviewed.

Studies on Sensitivity of Input Parameters During Building Design

Studying the thermal response of a building implies an understanding of how the various input parameters of a building interact with weather and with each other. The relationships involve both static variables (such as envelope U-factors and light and equipment densities) and dynamic variables (such as operational variables of building internal loads and control settings). The precise relationship between the simulation outputs and input design parameters cannot be expressed analytically because of the complex nature of their coupling. Therefore, sensitivity analysis and regression techniques are often employed. Insight provided by such analytical techniques will also be useful during the process of calibrating simulation inputs to monitored data.

The aim of sensitivity analysis is to determine or identify which input variables have a significant effect on the simulated output parameters (in our case, monthly utility bills that specify electric use, demand, and fuel use of the building) and then to quantify their relative importance. In mathematical terms, high-sensitivity elements thus identified can then be given more attention during the building design stage when alternative system design options are being investigated. As stated by Lam and Hui (1996), sensitivity theory could provide indication of the extent to which a particular building's thermal loads and energy consumption are responsive to changes in the coefficients of material properties, design of building envelope, selection and operation of HVAC systems, etc. The designer or the analyst can then spend their time on studying the effect of input parameters that matter most.

The general approach to determining sensitivity coefficients is summarized by Lam and Hui (1996): (1) formulate a base-case reference and its description, (2) study and break down the fac-

tors into basic parameters (parameterization), (3) identify parameters of interest and find their base-case values, (4) determine which simulation outputs are to be investigated and their practical implications, (5) introduce perturbations[3] to the selected parameters about their base-case values one at a time, (6) study the corresponding effects of the perturbation on the simulation outputs, and (7) determine the sensitivity coefficients for each parameter, if appropriate.

Sensitivity coefficients (also called *elasticity* in economics, as well as *influence coefficients*) are defined in various ways (Lam and Hui 1996): (a) as partial derivatives of the output variable with respect to the input variable, (b) as the base-case values to express the sensitivity in percentage change, (c) as the mean values to express the percentage change (this is similar to forward differencing and central differencing approaches used in numerical methods). Formulation (a) is used in comparative energy studies because the coefficient thus calculated can be used directly for error assessment. Formulations (b) and (c) have the advantage that the sensitivity coefficients are dimensionless. The results of several papers that performed sensitivity analyses and regression analysis on energy use in commercial buildings are briefly summarized below.

Sullivan et al. (1985) developed multiple linear regression equations that correlate the effect of various configuration parameters on building annual cooling and heating energy use and on peak cooling. They used DOE-2 to generate the basic energy use data for a large multi-story prototypical commercial building in five widely different geographic locations (Madison, WI; Lake Charles, LA; Seattle, WA; Washington, DC; El Paso, TX). Though correlating climatic variation was not the intent of this study, it allowed the authors to determine how the correlation structure was affected by climate. The building had multiple zones with one CAV system for each zone. One of the important tasks of the study was to identify the appropriate independent variables to use in the regression model. Configuration variables were mostly building physical characteristics that included orientation, wall and roof conductance, glazing properties of windows and skylights, and installed lighting power. Incremental effects of external shading (overhangs) and daylighting were also examined. Other relevant studies that have adopted the multiple linear approach to correlate the effect of different design parameters on building performance include those by Chou and Chang (1993) and Chou et al. (1993).

Corson (1992) studied the effect of various inputs for two buildings in the Pacific Northwest and found that the models were less sensitive to inputs related to building envelope and lighting and more sensitive to inputs related to occupancy, weather, air supply, systems, and plant.

Huang et al. (1993) studied the impact of variation in building conditions on energy use patterns so as to assess market potential for specialized applications. The study used a set of 481 prototypical commercial buildings in 20 US urban areas to develop a set of representative buildings in terms of physical characteristics, system configurations, and operational parameters. Simulation results of two building types (one large office of over 50,000 ft^2 and one hospital) using DOE-2 are presented in the form of sensitivity coefficient tables. The study also found that energy use is more influenced by the range and variability of internal operating conditions than the building envelope parameters.

Lam and Hui (1996) considered a 49,000 m^2 office building in Hong Kong with 40 stories. DOE-2 was used to study the effect of 60 input parameters categorized into three main groups: (1) building load, including window shading coefficient (SC), window-to-wall ratio (WR), space air temperature (AT), equipment density (EQ), lighting density (LL), and occupancy density (OC); (2) HVAC system, including outdoor airflow rate (OA), summer thermostat setpoint

[3] The use of the word "perturbation" is somewhat misleading. Strictly speaking, perturbation involves a *small* change, such as that associated with uncertainty of an input variable. Here, the changes are relatively large, on the order of 10% or more of the numerical value of the variable in question.

(TS), fan efficiency (FE), and fan static pressure (FS); and (3) primary plant inputs, including chilled-water supply temperature (CH) and chiller COP (CP).

About 400 simulations were performed with input parameters changed one at a time at ten different levels. Sensitivity coefficient plots for (1) annual electric consumption, (2) peak electric load, (3) peak heating and cooling loads, and (4) diurnal load profiles are shown along with regression fits to these plots, which were very good. Again, they found that the building envelope parameters are less influential than occupancy, weather, air supply, and system and plant parameters.

Motillo (2001) reports the results of a sensitivity analysis by building type using the DOE-2 software engine for five Canadian locations. Ten buildings were considered—three office buildings, one office-retail, two schools, one retail, two museums, and one multi-family residential building. Fourteen parameters were varied one at a time (over 100 simulations performed) to determine the annual energy change. Conversely to the other studies, he found that the thermal resistance of walls, roof, and fenestration had the most impact on annual energy consumption. Installed lighting power density, minimum outdoor airflow rate, pump type, efficiency of the heating equipment, and temperature setpoint schedules also had important effects. The parameters that had less impact were building orientation, thermal mass, service water heating equipment efficiency, supply airflow rate, average monthly ground temperature, and cooling equipment efficiency. The sensitivity results should be interpreted with care since they are specific to the Canadian climate and to the type of systems selected.

Reducing the Number of Simulation Runs

The number of runs or simulations to be performed, if an exhaustive grid spacing approach is adopted, increases exponentially with the number of parameters and the levels assumed. For example, if sensitivities of 12 input parameters at three different levels need to be determined, this requires $3^{12} = 531{,}441$ simulations to generate the complete sensitivity map. There seem to be three different ways by which this number can be reduced.

The *randomized approach* (Hui 1997) is akin to the Monte Carlo (MC) method, but instead of relatively small perturbations in the input data chosen according to a preselected probability distribution, the various inputs are randomly assigned numerical values within a physically consistent, prespecified range of minimum-maximum variation assumed to be normally distributed. Hui (1997) builds on the earlier paper by Lam and Hui (1996), where a sensitivity analysis was first done to identify 12 important parameters from an initial set of 60 variables. The 12 parameters consisted of 6 related to the building shell, 4 to the secondary systems, and 2 to the chiller. It was found that only 100 randomized simulations were adequate to generate enough information about the variation of annual electricity use and peak loads to correlate them with the 12 parameters with very good accuracy.

Experimental design methods from traditional statistics (Box et al. 1978) can provide ways to minimize the number of simulations. A conference paper by Hou et al. (1996) proposed a two-level factorial design approach to vary certain key input parameters of the DOE-2 program in order to develop building and HVAC system performance criteria for the Texas Building Energy Design Standard. Up to five factors were studied at two different Texas locations. The factorial design is of great value during the preliminary investigation stage when a large number of factors need to be examined superficially rather than a small number thoroughly.

Lattice methods, or *space-filling methods*, aim at testing the input space more uniformly (i.e., selecting a minimum number of combination sets of the input variables to perform simulations) yet allow multi-dimensional global optimization. Saporito et al. (2001) proposed a lattice method for global optimization (LMGO) for reducing the number of tests. The method was then applied to investigate the combined effect of thermal mass and other primary parameters

(mainly envelope related) on heating energy use in a UK office buildings using a detailed simulation program called APACHE. One case study illustrated how the effect of four variables, three with three levels and one with nine levels, can be reduced from 243 combination runs to 15 combinations. Another case study with 12 variables with five levels each was considered, and they showed how 5^{12} combinations can be reduced to 360 combinations. This method can generally be applied to optimization problems with a large number of combination tests and also for sensitivity studies. However, the paper points out that for sensitivity studies, the efficiency in terms of number of tests and computing time is probably very close to other classical methods such as the MC method.

APPENDIX B
SENSITIVITY ANALYSIS RELATED TO UNCERTAINTY

The concept of sensitivity analysis was discussed in Appendix A as it relates to the building design process. In that context, the effect of relatively large changes in the numerical values of the parameters on the simulated building performance was studied. Here, we are interested in relatively smaller changes in the parameters, as would be the case with uncertainty.

Lomas and Eppel (1992) point out that there are two types of sensitivities: (1) individual sensitivities that describe the influence of a single input variable (this is the case covered by Lam and Hui [1996], for example) and (2) total sensitivities due to variation in all the input data. Their paper compares three different sensitivity analyses methods: differential sensitivity analysis (DSA), Monte Carlo analysis (MCA), and stochastic sensitivity analysis (SSA) when applied to three different detailed simulation programs. The total sensitivities, in both hourly and daily average predictions, due to uncertainties in over 70 parameters were compared. It was concluded that DSA is appropriate to obtain sensitivities of predictions to individual input parameter uncertainties and that MCA can be used to obtain the total sensitivities in the predictions. The SSA method was found to require code modification, complex processing, and even more demanding computation effort, and it did not yield good results when applied to building energy simulation programs. The use of the SSA method was not recommended pending more work.

The individual sensitivity coefficient of the DSA method is the same as that described by Lam and Hui (1996). In order to compute total sensitivity, an average value of the sensitivity over the likely range of input parameter change is the most useful measure to estimate given the nonlinear relationship between input and output. The computational method by Judkoff et al. (1983) is advocated for small changes in the input data. Assuming that each input is normally distributed about the modal value, the combined or total influence on the predicted parameter Δp_{tot} is estimated from the quadrature sum of the influences due to each of the i inputs where each input is varied by the same amount (2.33 times its standard deviation if 99% confidence levels are chosen):

$$\Delta p_{tot} = \left(\sum_{i=1}^{I} \Delta p_i^2 \right)^{1/2} \tag{B1}$$

The total uncertainty given by Equation B1 is also at the 99% confidence level. This approach is strictly correct only if the sensitivity to each individual input is independent of the value of the other inputs. This is not true, of course, but this assumption is reasonable for *small* changes in the input data. However, during the calibration process, changes in inputs may be on the order of 10% or more, and whether this method of determining sensitivity coefficients is accurate for these cases remains to be seen. The two major drawbacks are (1) due to nonlinear input-output behavior, the number of numerical levels for each input variable needs to be large (for example, Lam and Hui [1996] chose to study ten different numerical levels uniformly spread over the

range of variability of each parameter) and (2) the number of simulations increase exponentially with the number of input parameters.

The MCA has been used widely in various scientific and engineering fields. In MCA, all the uncertain inputs must be assigned a definite probability distribution. For each simulation, one value is selected at random for each input based on its probability of occurrence. Numerous such input sequences are generated and simulations performed. Provided the number of runs is large, the simulation output values will be normally distributed irrespective of the probability distributions of the inputs. Though any nonlinearity between the inputs and output is accounted for, the accuracy of the results depends on the number of runs, and typically hundreds of runs may need to be performed. Further, the sensitivity of a single input cannot be gleaned from such simulations. Since detailed simulation programs, if run in batch mode, do not need long run times, this approach may be suitable for calibrated building energy simulation programs.

Macdonald and Strachan (2001) suggest a process for validation that entails adopting (1) structured validation using measured data and DSA and (2) unstructured validation using MCA. They argue that DSA was devised for practical experiments with relatively few (typically fewer than a dozen) measured and controllable inputs. When applied to simulations, which have hundreds of input parameters, the techniques become cumbersome and time consuming. On the other hand, MCA requires only about 60–80 simulation runs, after which only marginal gains in accuracy are gained. The study proposes uncertainty distributions for several of its inputs and suggests the use of MCA to study the effect of uncertainty propagation. Such a procedure has also been incorporated into a building thermal energy simulation program.

De Witt and Augenbroe (2002) address uncertainties in building performance evaluations and their potential on design decisions. The paper gives a brief overview of sources and techniques to determine the uncertainties in material properties and, more importantly, those stemming from model simplifications. The paper suggests a statistical screening technique to determine which sources have dominant effects on the outcome of the simulation. They suggest the use of the MC simulation technique as a crude first screen to determine which among 89 parameters are most critical. A sensitivity analysis using the factorial sampling method is used next, and finally a Bayesian analysis is performed to assess the consequences of each action being evaluated on the effect of indoor human comfort. The procedure is illustrated for a simple building envelope and considered parameters such as wind speed, indoor air distribution, and envelope material and heat transfer coefficients.