A Framework for Empirical Validation of Building Performance Simulation under Uncertainty

Qi Li[1], Godfried Augenbroe[1], Ralph Muehleisen[2]
[1]School of Architecture, Georgia Institute of Technology, Atlanta, GA, USA
[2]Argonne National Laboratory, Lemont, IL, USA

Abstract

This paper proposes an empirical validation framework to validate building performance simulation tools under uncertainty. It presents a case study of a future experiment in a controlled environment with detailed measurement. At this stage synthetic measurement data were generated to demonstrate the framework and preliminarily assess the experiment. The simulation and analysis result highlights the importance of explicit uncertainty quantification especially by means of detailed measurements. It also demonstrates the benefit of probabilistic agreement criteria. Future work on actual experiment is expected to extend this framework for practical use.

Introduction

While building performance simulation (BPS) has been widely used in both academic and industrial applications, its capability of predicting actual performance and informing associated decision-making lacks full understanding. This may lead to its underutilization and lost opportunities in practice due to practitioners' lack of confidence in BPS tools.

To address this and other lingering concerns about the performance gap between predicted and actual energy performance, the U.S. Department of Energy (DOE) initiated the research project "Validation and Uncertainty Characterization for Energy Simulation: Unlocking Opportunities in Energy Efficiency". It aims to overcome the barrier of adopting building performance simulation in practice by reducing the uncertainty in performance prediction, by means of both identifying errors and inadequate assumptions inside simulation engines and systematically characterizing associated uncertainties in component and system models. To this end, this study proposes a framework to validate BPS tools by empirical validation based in part on explicit uncertainty quantification, such as to help identify issues to be addressed in future updates of the simulation software.

While the rest of the paper describes application of the framework specifically to EnergyPlus (Crawley et al. 2000) models, the purpose is to develop a framework that can be applied to any dynamic simulation model with hopes that the framework will be deemed robust enough to become an integral part of future validation standards.

Background

Empirical validation, alike analytical verification and inter-program comparison, has been widely recognized as a useful method for validating a simulation program (Judkoff and Neymark, 2006). However, its complexity and high time and labor cost have resulted in only a few related studies in the literature that either focused on single-room climate chamber or test cells experiments. This includes IEA Annex 21 (Lomas et al. 1997), IEA Task 22 (Palomo del Barrio and Guyon, 2003; 2004), and PASSYS project (Jensen, 1993; Clarke et al., 1993). It also applies to realistic full-scale residential building experiments as done in IEA Annex 58 (Strachan et al., 2015; 2016). While the procedures and experimental design have been well established, the theoretical validation framework especially with respect to handling model and experiment uncertainty and establishing test performance criteria is relatively less studied. Palomo et al. (1991) suggested using residual analysis to assess simulation performance, identify potential causes, and improve model prediction. The authors proposed various statistical metrics to quantitatively characterize the discrepancy and inform model improvement. Palomo Del Barrio and Guyon (2003, 2004) proposed a two-step empirical model validation methodology that includes checking model validity and diagnosis. The first step relies on residual analysis and comparison between model outputs uncertainty bands and measurements uncertainty intervals. Based on linear assumption, the second step uses local sensitivity analysis to identify parameters influential to model discrepancy, and uses optimization techniques to search for parameter values that reduce the discrepancy. Strachan et al. (2016) used the absolute difference and Pearson correlation coefficient to assess the magnitude and profile fitness of model prediction.

This study differs from the existing literature in the following sense. First, instead of solely inverse parameter identification, which could potentially bias parameter values to subsume model error, this framework relies on explicit forward quantification of parameter uncertainty through detailed experiment monitoring, such that parameter error and model form error can be assessed separately. Second, it proposes the use of probabilistic discrepancy measures to characterize the agreement between a single realization and an underlying distribution such as to assess experiment adequacy.

https://doi.org/10.26868/25222708.2017.466

Framework overview

Conventional model validation method is based on quantifying the discrepancy between a model prediction and corresponding measurement in the field experiment, and compare this discrepancy with a "standard" threshold. Because of the associated uncertainty in both the model and the experiment, a poorly designed experiment with huge uncertainties may not be able to differentiate valid and invalid models, as they all exhibit large discrepancy with the measurement due to those uncertainties. Therefore, this study proposes to construct an "internal" (i.e. used only by the experimenter) probabilistic model prediction that fully considers all sorts of uncertainty, and use its discrepancy with the measurement, presumably the "best guess", as the threshold instead to evaluate other models, i.e. "external" models. In addition, the discrepancy of this "best guess" itself becomes an indicator of the adequacy of the empirical validation experiment. If a huge discrepancy between the internal model and the measurement is present, one cannot expect this experiment to be able to detect and reject poor external models.

From this perspective, the proposed empirical validation framework includes the following steps:

1. Given the available facility information and experiment setup, create a corresponding internal energy model and identify and quantify the uncertainty of all model parameters, informed by empirical knowledge as well as monitored data.
2. Generate a probabilistic distribution of the measured outcome by non-intrusively propagating parameter uncertainty through experimental design, random sampling, and repeated simulations.
3. Assess the validity of the experiment by checking this distribution's agreement with actual measurements based on introduced probabilistic discrepancy measures. Large discrepancy indicates the inadequacy of the experiment, along with its information, setup, and measurement, to validate an external simulation model. Measures to constrain associated uncertainty and thus reduce the discrepancy, for example an experiment repeat with more detailed measurement, should then be taken to improve the experiment.
4. Once validated, the level of agreement will serve to assess and validate external simulation models.

The rest of the paper deals with the description of two yet-to-perform empirical validation experiments. Using EnergyPlus 8.5.0 model as an example, it explains in detail the methods to quantify the associated parameter uncertainties of a detailed dynamic simulation model based on both empirical knowledge and particular experimental measurements. It illustrates how uncertainty and sensitivity analyses generate the distribution by using synthetic measurements. It also identifies influential sources of potential discrepancies, quantified by proposed probabilistic discrepancy measures. It concludes with a discussion of the results with respect to implementation implications.

Validation experiment

Lawrence Berkeley National Laboratory (LBNL) is going to perform a series of experiments to benchmark current building simulation practice for performance simulation of a single zone conventional mixing system. Two of these future experiments, the idealized low mass, no-window tests of a convection electric heating (co-heating, I-430-E) and air handling unit cooling (AHU cooling, I-430-C) that resemble BESTEST (Judkoff et al. 2010) #430 tests, are used as a case study in this paper.

Experiment facility and setup

The 20×30×12ft. test cell #1B is the experiment single zone used for the targeted experiments. It has two adjacent test cells #1A and #2A on the east and west respectively. It also has a mechanical room and an electrical room adjacent to the north. Its south exterior wall and window are fully configurable; the rest of the envelope, including roof and all partition and exterior walls, are equipped with R-80+ insulation. The experiments will add 4" polyisocyanurate board, of approximately R-23 insulation, to the window reveal to emulate no-window condition. The test cell has a radiant slab with embedded hot water tubes, whose temperature will maintain to be equal to the anticipated room temperature to minimize heat transfer in addition to extra insulation. Idealized internal loads from manikins only apply to the AHU cooling experiment.

Figure 1: FLEXLAB test cell #1A (left), #1B (middle), and #2A (right, partially shown)

A brief explanation of the experiment plan upon which this study is based is presented here, which may be subject to changes in final realization. The co-heating test will use multiple electric heaters to maintain the room temperature to be around 10 °C above the ambient air temperature over a three-day experiment period; the AHU cooling test will use the cooling coil in the dedicated on-site AHU to maintain an indoor-outdoor temperature difference of around 5 °C over four days. Interior fans will be utilized to ensure the room air is well mixed. Monitoring of the experiment conditions and outcomes includes:

- Room air temperature at various locations.
- Interior surface temperature and heat flux of all envelope components at various locations.

- CO_2 tracer gas injection rate and room concentration for infiltration measurement.
- Heater power in the co-heating test; flow rate and inlet/outlet temperature of AHU chilled water and supply air, fan power, and internal load in AHU cooling test.
- Meteorological conditions, including on-site dry bulb and dew point temperature, global and diffusive irradiance, wind speed and direction, and potentially ground reflectance.

Dynamic simulation model

A model in EnergyPlus 8.5.0 is created according to the architectural drawings and experiment setup (Figure 1). To reduce external uncertainties, all adjacent rooms are neglected, and the boundary conditions are specified by measured surface temperature on the other side of related partition walls. Complexity of construction in terms of corrugation, studs, box-in columns are neglected. Shading from adjacent test cells are modelled by shading surfaces accordingly.

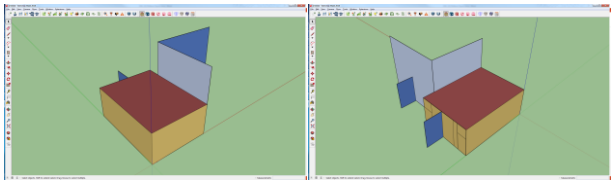

Figure 2: FLEXLAB cell model in EnergyPlus 8.5.0

The radiant slab is modelled in the same way as the partition walls by predefined boundary surface temperature schedules. Explicit modelling of the ground, the structural slab and the tube-embedded radiant slab is therefore neglected; only the insulation panel, below which 9 surface temperature sensors are to be installed, is modelled as the sole floor construction. Both the portable heaters and the AHU are modelled according to the specifications, except that the actual cooling coil operating on 30% glycol is simplified to a normal chilled water cooling coil because of limitation in the EnergyPlus water coil model.

Parameter uncertainty quantification

This section explains the parameter uncertainty quantification methods in this study that will be applied to the actual tests. Detailed experiment information and measurements to be expected in the tests are used to quantify case-specific uncertainties. Uncertainties of the remaining parameters are adopted from a generic uncertainty quantification repository (Wang 2016) that is based on empirical knowledge, existing literature, and previous experience.

Material properties

Uncertainty information of envelope material properties comes primarily from Macdonald (2002). All other parameter uncertainties are assumed based on previous experience and knowledge of similar properties of other materials. Bounded normal distribution are assumed for each property parameter, whose standard deviation (std.) are summarized in Table 1.

System equipment characteristics

Uncertainty of some fan characteristics can be found in the literature (AHRI 430 2009; ANSI/AHRI 2011; Griffith et al. 2008). Uncertainty of remaining component properties and heating coil capacity are assumed based on previous experience. Since EnergyPlus uses design conditions to specify water cooling coil characteristics to which the simulation robustness is very sensitive, the analysis neglects its uncertainties. Table 1 summarizes the associate uncertainties as well.

Table 1: Model parameter uncertainties

Model parameter	Std.	Unit
Material properties		
Opaque material		
Conductivity[1]	5%	W/m/K
Density[1]	1%	kg/m3
Specific heat[1]	12.25%	J/kg/K
Thermal absorptance[1]	2%	-
Solar absorptance[1]	7%	-
Visible absorptance	7%	-
Air gap		
Thermal resistance	5%	m2/K/W
Glazing		
Solar transmittance	1%	-
Front side solar reflectance	1%	-
Back side solar reflectance	1%	-
Visible transmittance	1%	-
Front side visible reflectance	1%	-
Back side visible reflectance	1%	-
Infrared transmittance	1%	-
Front side infrared emissivity	1%	-
Back side infrared emissivity	1%	-
Conductivity[1]	5%	W/m/K
Dirt correction factor	10%	-
System component: Electric Heater		
Heating coil capacity	1%	W
Fan total efficiency[2]	10%	-
Fan pressure rise	5%	Pa
Fan motor efficiency[2]	5%	-
System component: Air Handling Unit		
Heating coil rated capacity	1%	W
Fan total efficiency[2]	7.5%	-
Fan pressure rise	5%	Pa
Fan motor efficiency[2]	5%	-

Source: 1. MacDonald (2002); 2. AHRI 430 (2009); Griffith et al. (2008); 3. ANSI/AHRI 2011; Griffith et al. (2008)

Convective heat transfer coefficient

In this study, uncertainty quantification of convective heat transfer coefficient for interior and exterior surfaces of building envelope employs the approach proposed in Sun (2014). This approach builds upon the DOE-2 convective heat transfer model and quantifies the uncertainty associated with model coefficients concerning natural and forced convection, and the bivariate joint distributions are derived using meta-analysis for vertical wall, floor and

ceiling surface individually. More details can be found in Sun (2014).

Sensor error

Calibration of sensors is to be performed before the experiment, upon which the associated uncertainty quantification is based. In this experiment it is assumed that the reference sensor reading has no bias and same variance from the manufacturer (Table 2), and the to-be-calibrated sensor reading has an individual constant bias and random error. Let v^t denotes the true physical value to be measured at time step t, $t = 1,2,\dots k$; δ_{c^i} denotes the bias of the ith to-be-calibrated sensor, $i = 1,2,\dots n$; ε_r and ε_{c^i} denote the random error of reference and to-be-calibrated sensor reading respectively, then for the reference sensor,

$$v_r^{t\prime} = v^t + \varepsilon_r, \varepsilon_r \sim \mathcal{N}(0, \sigma_M^2) \quad (1)$$

In which σ_M^2 can be estimated from manufacturer specifications, as shown in Table 4, by assuming the range equals three times of the standard deviation. For the to-be-calibrated sensor,

$$v_{c^i}^{t\prime} = v^t + \varepsilon_{c^i}, \varepsilon_{c^i} \sim \mathcal{N}\left(\delta_{c^i}, \sigma_{c^i}^2\right) \quad (2)$$

Therefore, their difference at each time step has:

$$\Delta^t = v_{c^i}^{t\prime} - v_r^{t\prime} = \varepsilon_{c^i} - \varepsilon_r \sim \mathcal{N}\left(\delta_{c^i}, \sigma_{c^i}^2 + \sigma_M^2\right) \quad (3)$$

which gives the estimator of δ_{c^i} and $\sigma_{c^i}^2$ as:

$$\hat{\delta}_{c^i} = \bar{\Delta}^t, \hat{\sigma}_{c^i}^2 = \frac{1}{k-1}\sum_{t=1}^{k}(\Delta^t - \bar{\Delta}^t)^2 - \sigma_M^2 \quad (4)$$

Surface boundary conditions

For surfaces with adjacent boundary conditions in the experiment, the surface temperature sensor readings will be used to construct a time-varying boundary condition that simplifies the model and reduces the control volume. For an individual sensor, it is believed that the uncertainty comes only from time independent sensor errors. First of all, for each surface temperature sensor reading $T_{c^i}^{t\prime}$ it has:

$$T_{c^i}^{t\prime} = T^t + \delta_{c^i} + \varepsilon_{c^i} \sim \mathcal{N}\left(T^t + \delta_{c^i}, \sigma_{c^i}^2 + \sigma_M^2\right) \quad (5)$$

where T^t is the true temperature. Then for the area-average mean value, which acknowledges the non-uniform temperature distribution on the surface, it gives:

$$\hat{T}^t = \frac{1}{A}\sum_{i=1}^{n} A_i\left(T_{c^i}^{t\prime} - \hat{\delta}_{c^i}\right)$$
$$Var\left(\hat{T}^t\right) = \frac{1}{A^2}\sum_{i=1}^{n} A_i^2\left(\sigma_{c^i}^2 + \sigma_M^2\right) \quad (6)$$

where A_i is the representative area of each sensor reading and unknown parameters are to be replaced by estimators from sensor calibration. Similarly, in case of comparing hourly simulation and measurement only, assume that each hour the temperature remains constant and equals the mean value, so to aggregate sub-hourly readings within each hour t_s:

$$\hat{T}^{t_s} = \frac{1}{s}\sum_{j=1}^{s}\hat{T}^j, Var\left(\hat{T}^{t_s}\right) = \frac{1}{s^2}\sum_{j=1}^{s} Var\left(\hat{T}^j\right) \quad (7)$$

Equation (6) and (7) clearly show that aggregation over space and time reduces the uncertainty/variance. For uncertainty propagation, a random sequence of surface temperatures can be obtained by independent sampling at each time step, and a Monte Carlo sample using Latin Hypercube design (LHD) of random sequences will generate a sample of the uncertain surface temperature.

System control sensors

Because of lack of information, it is assumed that the system control sensors, including thermostat, AHU supply air temperature, and chilled and hot water temperature, have the same sensor error as those used in the measurement and are also implemented as random sequences.

Meteorological conditions

The local meteorological condition, as measured by an on-site station, is to be used as the simulation inputs. Accordingly, the sensor error is considered as the only source of uncertainty. Uncertainty of ground reflectance, on the other hand, is assumed based on experience because of lack of information. Table 2 also summarizes the associated uncertainties.

Table 2: Sensor error

Sensor type	Manufacturer error	Unit
Test cell conditions		
Surface temperature	±0.05	°C
Air temperature	±0.05	°C
Water temperature	±0.03	°C
Water flow rate	±0.41%	m3/s
Meteorological conditions		
Dry bulb temperature	±0.05	°C
Dew point temperature	±0.2	°C
Global irradiance	±5%	W/m2
Diffuse irradiance	±5%	W/m2
Wind speed	±0.1	m/s
Wind direction	±1	°
Ground reflectance	±0.1	-

Effective leakage area

For constant injection tracer gas method, ASTM Standard E741-11 (ASTM 2011) gives the formula to calculate average infiltration flow rate Q and associated error s_Q:

$$Q = Q_{tracer}\frac{1}{n}\sum_{j=1}^{n}\frac{1}{C_j} - \frac{V_{zone}}{t_n - t_1} ln\left[\frac{C_n}{C_1}\right] \quad (8)$$

$$\frac{s_Q^2}{Q^2} \approx \frac{s_{Q_{tracer}}^2}{Q_{tracer}^2} + \frac{s_C^2}{\bar{C}^2}\left(\frac{Var(C)}{s_C^2}\right)^2 + \frac{2{V_{zone}}^2}{(t_n - t_1)^2 Q^2} \quad (9)$$

where Q_{tracer} is the tracer gas injection flow rate, C_j is the tracer gas concentration at time step $t_j, j = 1,2,\dots n$, V_{zone} is the room volume, and s_* are the corresponding errors that can be estimated from time series readings. The

LBNL infiltration model is then used to translate the infiltration rate into effective leakage area:

$$ELA = \frac{10000 \cdot Q}{s}, s = \sqrt{C_s \cdot \Delta T + C_w \cdot v^2} \quad (10)$$

where uncertainty of ΔT and v come from weather station sensor error and aggregation over the entire experiment period using methods shown above. Uncertainty of C_s and C_w are neglected because of their relative insignificance. Therefore, a bootstrap sample of ELA can be obtained by independently sampling on Q, ΔT and v and calculate the ELA correspondingly.

Internal load

In this experiment the internal load comes from both manikins and interior mixing fans. Their uncertainties come only from manufacturer specifications, assuming no heat loss during energy conversion from electricity to heat. Uncertainty in fraction of radiation is obtained from specification or otherwise assumed.

Outcome measurement

Similarly, for output measurement to be used to validate simulation results, the uncertainty from sensor error and aggregation can be obtained in the same way as shown above in Equation (6), except using volume-weighted average instead. In this study, the hourly mean room air temperature (referred to as room temperature hereafter), heating energy of electrical heater, and cooling energy cooling coil are used as the outcome of interest, whose respective measurement error is either derived from individual sensor errors or assumed in absence of information. Table 3 summarizes the standard deviation of their respective normal distributions.

Table 3: Observation error

Measured outcome	Std.	Unit
Room temperature	0.005	°C
Heating energy	0.1%	kJ
Cooling energy	0.5%	kJ

Uncertainty and sensitivity analyses

While the above uncertainty quantification methods will be applied to the actual experiments when they are performed, in order to demonstrate the framework in this paper, "virtual" experiments using synthetic sensor data and empirical knowledge have been performed to quantify and propagate the associated uncertainties.

Uncertainty allocation and propagation

Uncertainty estimates for generic material properties and system equipment are straightforwardly added to the EnergyPlus model. Uncertainty of convection heat transfer coefficient is implemented by using objects *SurfaceConvectionAlgorithm:Inside:UserCurve*, *SurfaceConvectionAlgorithm:Outside:UserCurve* and *Curve:Linear/Exponent* to quantify uncertainties through curve coefficients. Estimates are made for the uncertainty quantification based on detailed measurement as available. Error ranges from manufacturer specifications are used both to define the range from which a synthetic bias is uniformly sampled for a particular sensor, and to derive the random error variance. Surface boundary conditions are defined by averaging over 9 synthetic temperature sensor readings with equal representative area on each surface; the random error variance is therefore 1/3 of a single sensor. These boundary conditions are then imported to the model by object *SurfaceProperty:OtherSideCoefficients*. A LHD Monte Carlo sample of 400 sequences is generated by sampling for each time step independently, which is then imported through object *Schedule:File* into the model during the simulation. Random sequences of system settings are also generated and added similarly.

This case study uses TMY weather file from the nearby Oakland International Airport. The ground reflectance is assumed to be uniformly distributed from 0.1 to 0.4. All the other associated uncertainties are neglected for simplicity. Distribution of effective leakage area is directly assumed according to the generic UQ repository: a lognormal distribution with a mean of -1.6551 and a standard deviation of 0.8767. Likewise, the electricity unit heater is auto-sized, and the mixing fans are modelled also as internal load with assumed power.

The two tests are assumed to be performed on 01/18-01/20 and 07/18-07/20 respectively, with a thermostat setting of 30°C for heating and 18°C for cooling. Parameters identified as random variables are sampled from their probability density functions. Boundary conditions identified by random sequences are determined by sampling on a random variable uniformly distributed from 0 to 1, whose value corresponds to a particular sequence in the previously constructed sequence sample that is to be used in the simulation. Only group uncertainty is considered for simplicity, i.e. a single sample point value is assigned to different instances, so for example different walls have the same convection heat transfer curve coefficients within each sample point. Uncertainty in outcome measurement is represented by random sampling on the distribution conditioned on model parameter uncertainties, i.e. in each Monte Carlo sample point a random outcome sequence is generated by sampling within its uncertainty band. This in the end results in a LHD sample of 2000 points of 159 parameters.

Simulation result

This section shows the outcome of the simulation of the constructed EnergyPlus model for the controlled experiment with added parameter uncertainty. Results in terms of the empirical 95% confidence uncertainty band can be found in Figure 3. It clearly shows that the thermostat setpoint is well maintained in the co-heating test, whereas the room temperature has large variations during the AHU cooling test and is constantly below the cooling setpoint. This could be because the cooling load is very small due to the mild local weather, the blocked window and the assumed internal load. Since the system has no reheat coil, the cool air is continuously supplied to the zone at minimum flow rate and therefore the room temperature is in principle free floating. In the meanwhile, large variation is observed in both heating and cooling

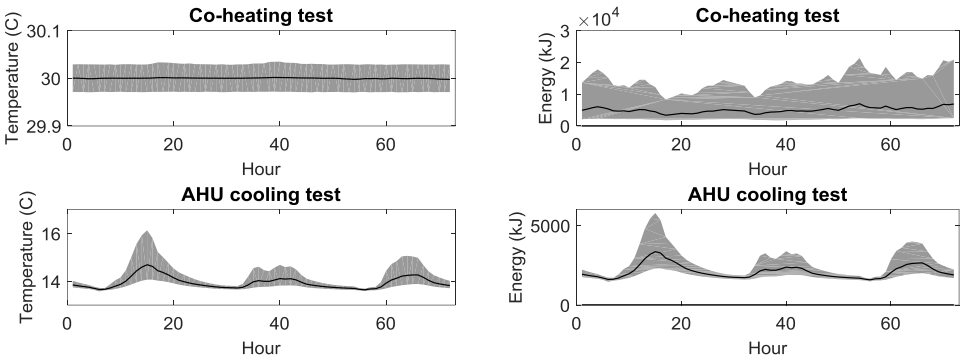

Figure 3: Room temperature (left) and heating/cooling energy (right) with empirical 95% confidence intervals (shade)

energy, suggesting the potential need for extra measurements to reduce associated uncertainties.

Sensitivity analysis

A sensitivity analysis using the Morris method (Morris 1991) is performed to identify important parameters responsible for output variations. A detailed explanation of the procedure and underlying logic can be found in Menberg et al (2016). In this case study, a 10-level design of the 159 parameters is constructed, resulting in a sample of size 1600. As the Morris method handles only univariate output, principal component analysis (PCA, Hotelling, 1933) is used to aggregate the simulation time series output. It represents each original time series output $\boldsymbol{Y}_T^i, , i = 1,2, \dots n$, as a linear combination of a particular set of orthogonal unit vectors $e_j, j = 1,2, \dots T$:

$$\boldsymbol{Y}_T^i = \sum_{j=1}^{T} w_j^i e_j \qquad (11)$$

These unit vectors, i.e. principal components, are chosen to maximize the variation of the linear coefficients w_j^i, i.e. scores. Therefore, a reduced representation of the original output, using only a few principal components and their associated scores, maintains the capability to distinguish individual output without much loss of information. This study only chooses the first principal component of each temporal output, and the scores w_1^i is used as the aggregated univariate output.

$$w_1^i = \boldsymbol{Y}_T^i {e_1}^{-1} \qquad (12)$$

Percentage of variations explained by the first principal component of each output is shown in Table 4, indicating an overall adequacy of the first principal component in reflecting the variations due to parameter uncertainty and thus in identifying the influential parameters. The only exception is room temperature in co-heating test as it is well maintained and therefore sensitive to noisy disturbance in the heater control temperature sensor.

The absolute mean μ^*, i.e. the expected change in output due to variation in each parameter, is used to rank parameters. The standard deviation of those changes of an individual parameter, σ, reflects the magnitude of its non-linear effects and interactions with other parameters. Figure 4 plots the $\mu^* - \sigma$ plots of each output and identifies the influential parameters, in which parameters appearing on the right have large overall impact on the output and those appearing on the top have large non-linear effects and interactions with other parameters.

Table 4: Percentage of variation explained by the first principal component

Measured outcome	Percentage
Room temperature, co-heating test	72.10%
Heating energy, co-heating test	99.90%
Room temperature, AHU cooling test	93.76%
Cooling energy, AHU cooling test	93.61%

Note that each convection curve consists of two coefficients that are sampled from a joint distribution driven by two independent random variables, so the result here is for basic understanding of important mechanisms, and a group sensitivity analysis such as group lasso (Yuan and Lin, 2006) would be more appropriate for a quantitative interpretation. Nevertheless, the results clearly show that effective leakage area and envelope convection, whose uncertainties are assumed based on generic UQ repository, contribute to most of the variations. This underscores the importance of additional measurement in deriving case-specific uncertainties (which one would hope would be reduced) especially if there are large variations in the energy outcomes.

Probabilistic discrepancy measures

In the proposed empirical validation methodology, using the measured outcome from the experiment to assess the distribution of the internal simulation is based on comparing a single outcome (e.g. from an external simulation test) with a distribution resulting from an internal model. In order to validate external models, the internally generated distribution should satisfy the following two criteria: 1) accuracy: it agrees with the measured outcome well enough; 2) sharpness: its range, i.e. output uncertainty is small enough to be able to accept or reject other models with enough confidence. Most empirical validation literature concerning modelling and measurement uncertainty assess the agreement simply by comparing two uncertainty bands. Palomo Del Barrio and Guyon (2003) attempts to satisfy the criteria by the use of a heuristic bounding algorithm, which however ignores probable abnormal probability density caused by model

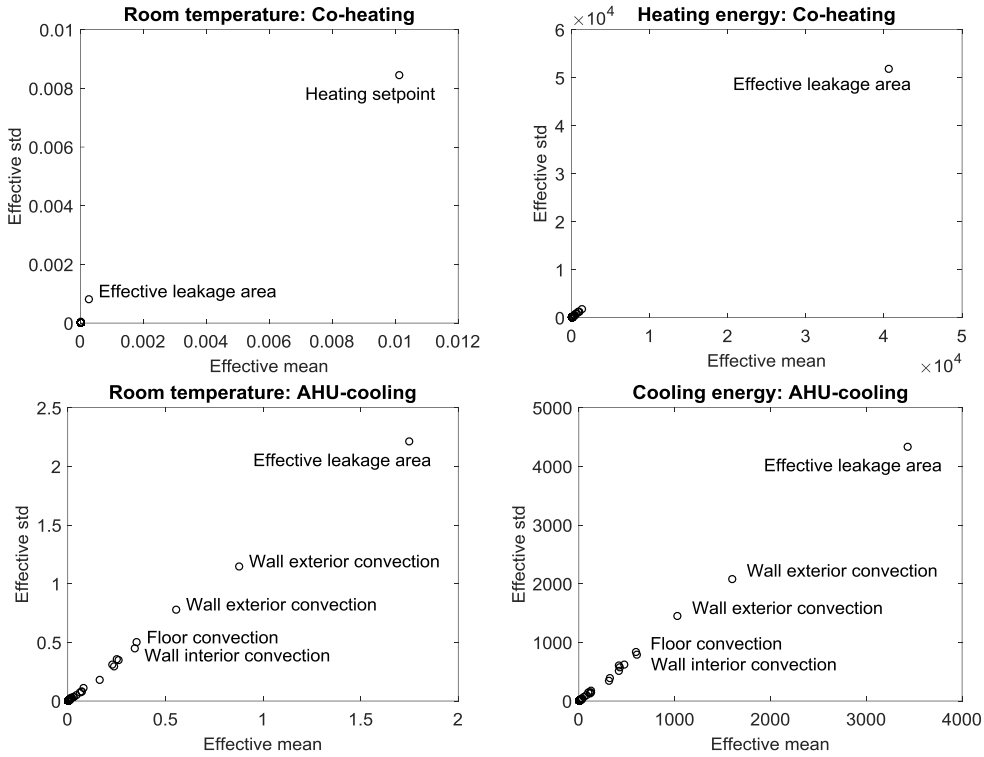

Figure 4: Sensitivity analysis result: $\mu^* - \sigma$ plots

uncertainties and their complex non-linear effects and interactions. As an alternative, this study proposes two probabilistic discrepancy measures, i.e. the probability of an instance according to a distribution, and continuous rank probability score (CRPS), as two alternative quantitative characterizations of model agreements.

Instance probability

As a probabilistic discrepancy measure, the instance probability represents directly the probability that an instance happens according to a certain probabilistic distribution. If the probability is small, this indicates either a large disagreement between the instance and the distribution, or a large dispersion of the distribution; both correspond to the aforementioned criteria.

To calculate the instance probability, a kernel density estimator is constructed from the discrete output of Monte Carlo simulation such as to estimate the continuous probability density function (PDF):

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x_i}) \qquad (13)$$

where $\boldsymbol{x}$ is a d–dimension instantiation to be evaluated and $\boldsymbol{x_i}$, $i = 1,2,\dots n$ are the Monte Carlo simulation outputs, $K_h(\cdot)$ is the kernel function. For commonly used Gaussian kernel, it becomes:

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n2\pi^{\frac{nd}{2}}|\boldsymbol{H}|^{\frac{1}{2}}} P(\boldsymbol{x})$$
$$P(\boldsymbol{x}) = \sum_{i=1}^{n} exp\left\{-\frac{(\boldsymbol{x} - \boldsymbol{x_i})^T \boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{x_i})}{2}\right\} \qquad (14)$$

where $\boldsymbol{H}$ is the bandwidth matrix and an optimal can be found for Gaussian kernels.

Continuous rank probability score

The potential of continuous rank probability score to assess prediction agreement receives particular attention in forecast verification (Gneiting and Raftery, 2007) and also in evaluating building performance predictions (Sun, 2014; Li et al., 2016). One can calculate the score of a Monte Carlo simulation based probabilistic prediction by:

$$CRPS(F, y) = E_F|Y - y| - \frac{1}{2}E_F|Y - Y'| \qquad (15)$$

where F is the predictive distribution of random variable Y represented by the sample set in the form of cumulative distribution function (CDF), y is the single observation, E_F is the expectation over F, Y' is an independent random variable with identical distribution as Y, obtained by random permutations of the sample set F. Having the same unit as the original output, a large CRPS value indicates a large discrepancy between the predictive distribution and the single observation. Furthermore, the representation shows that the CRPS generalizes the absolute error, to which it reduces if F is a point forecast. Figure 5 shows the CRPS of a normal distribution prediction with different biases and standard deviations. It clearly shows that CRPS increases when the probabilistic prediction is either too far from observation or too dispersed, both also correspond to the criteria. One can refer to Gneiting and Raftery (2007) for more technical details.

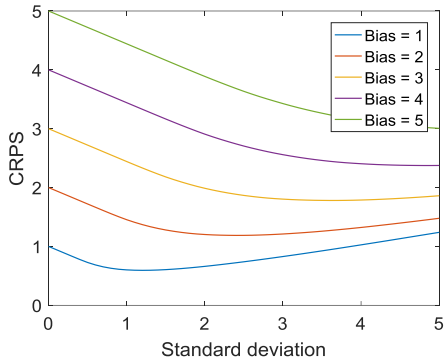

Figure 5: Illustration of CRPS with normal distribution

Exploratory testing

Without actual experiment data being available at this time, this paper uses synthetic instances and distributions to test the feasibility of this empirical validation methodology, as well as a preliminary assessment of the experiment's adequacy. The mean sequence of each simulation output forms the synthetic instance data, whereas the probabilistic distribution from previous Monte Carlo simulation output, with either a bias of ±5%, ±10%, and 15% of the mean or a dispersion with a standard deviation of 50%, 75%, 125%, and 150% of the original, forms the group of probabilistic distributions subject to testing of accuracy and sharpness respectively.

Regarding instance probability, as high dimensional kernel density function is computationally prohibitive, the previous first principal component of each output is used as the output. This paper considers those output separately for convenient interpretation, so a univariate version of Equation (14) is applied to each output. In contrast, as CRPS generalizes the absolute error and thus has the same unit, a simple time average value is used for each output and therefore does not involve principal components.

Result and discussion

Testing result of instance probability is shown in Figure 6. It tends to penalize distributions with large bias or dispersion, although probably because of skewness the mean sequence does not necessarily possess the largest probability as in the case of heating and cooling energy. Result of CRPS in Figure 7 shares similar pattern with respect to bias and dispersion. However, it is noteworthy that it has less asymmetry toward bias in AHU cooling test. One possible reason is that instance probability is PDF based whereas CRPS is CDF based, so the former is more sensitive to small probability profile fluctuations than the latter. Another possible reason is that the continuous PDF to calculate instance probability is estimated from a discrete Monte Carlo sample with a finite sample size, so it is possible to differ from the true PDF, and perhaps require a larger sample size to achieve better convergence. Nevertheless, the possible occurrence of skewness suggests the inappropriate common impression that closer to the center leads to better agreement, and underscores the importance of explicit uncertainty quantification.

In terms of experiment adequacy, using the probability densities at 3 and 1 standard deviation away from the mean of a standard normal distribution, i.e. 0.0044, and 0.2420 respectively, as two reference values, the result of instance probability clearly shows that the small variation of room temperature in the co-heating test causes a large sensitivity to bias and insensitivity to dispersion, making it difficult for an external instance (e.g. a generated simulation outcome) to be deemed consistent. On the other hand, a large variation in heating and cooling energy leads to an overall small instance probability, which also poses difficulties for validating instances and perhaps require further reducing uncertainty. Only the room temperature in the AHU cooling test seems to be adequate for validation purpose as it can properly detect the bias and dispersion variation with a reasonable instance probability. In contrast to instance probability, CRPS has intuitive physical meanings and therefore is an ideal metric to assess external instances with the result from real experiment for each individual type of output. However, this feature also makes it difficult to establish a universal threshold in validating external instances, as well as to compare different outputs in terms of their respective validation adequacy.

In summary, while both instance probability and CRPS in principle reflect accuracy and sharpness criteria in validating instances, their respective drawbacks warrant further study on real measurement data and practical situations in order to develop an appropriate empirical validation process. At the same time, the testing result tentatively highlights the potential inadequacy of room temperature as a measured outcome in co-heating tests. It also reiterates the importance of further measures on reducing uncertainties in heating and cooling energy output. Further study is expected to consolidate the above observations.

Conclusion

This paper proposes an empirical validation framework to validate building performance simulation tools against controlled experiments. A case study of a hypothetical experiment with controlled environment and detailed measurement is presented to demonstrate this approach. First the controlled experiment set up is captured in a detailed simulation model with added uncertainties in model parameters that cannot be known precisely. Uncertainty and sensitivity analyses of the simulation results are performed with quantified uncertainties and synthetic experiment data. Principal component analysis is then used to aggregate time series data into reduced dimensions. As the experiment is not yet conducted, synthetic measurement data is generated for the ensuing confidence analysis. In this analysis the simulated output is compared to the measurement data to determine the validity of the experiment for the benchmarking of external simulation models.

The simulation and analysis result highlights the importance of rigorous and case-specific uncertainty quantification for empirical validation. It demonstrates the potential of two probabilistic discrepancy measures in

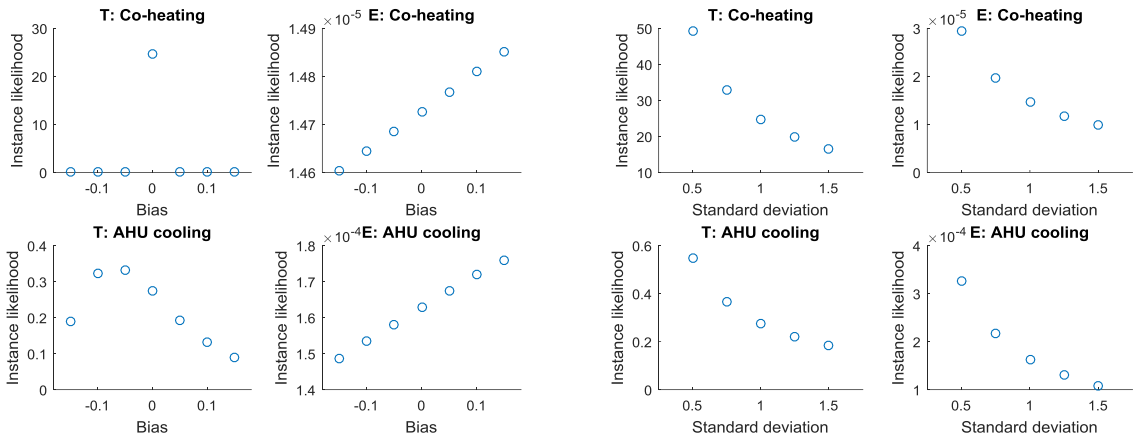

Figure 6: Testing result of instance probability on accuracy (left) and sharpness (right) (T: room temperature; E: heating/cooling energy, same hereafter)

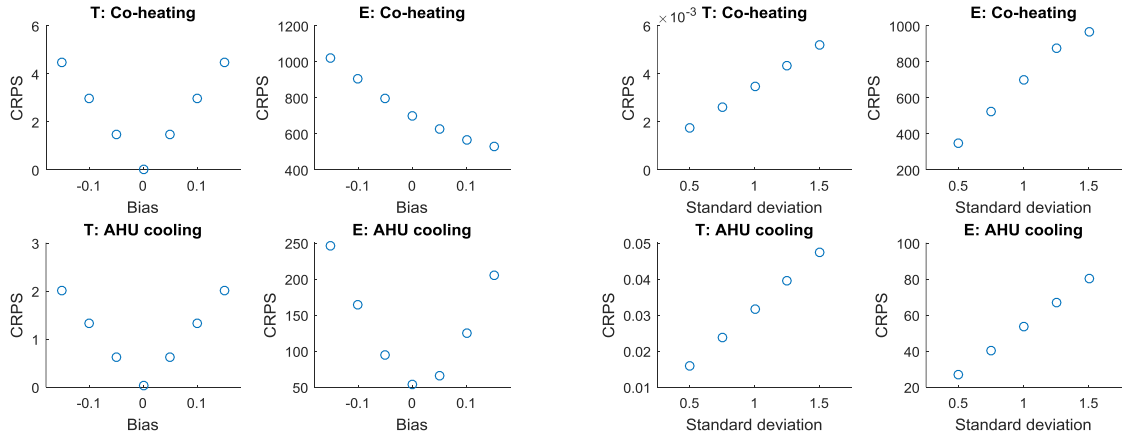

Figure 7: Testing result of CRPS on accuracy (left) and sharpness (right)

assessing experiment validity. In addition, as a preliminary assessment of the validation experiment, this paper suggests special attention should be paid to constrain uncertainties associated with infiltration and envelop convective heat transfer in an effort to improve its validation adequacy.

It is expected that the modelling approaches and uncertainty quantification methods used in the case study could inform similar practices that try to compare detailed measurement data with simulation predictions. It is also expected that, after a more thorough study on different experiments and models with actual measurement data being involved, explicit quantitative validation criteria can be established to appropriately evaluate the performance of building simulation tools for this and other empirical validation experiments.

Acknowledgement

This work was supported by the U.S. Department of Energy under Contract No. DE-AC02-06CH11357.

References

AHRI Standard 430 (2009). Performance Rating of Central Station Air-Handling Units, Air-Conditioning, Heating, and Refrigeration Institute, Arlington, VA.

ANSI/AHRI Standard 1211 (2011). Performance Rating of Variable Frequency Drives, American National Standards Institute and Air-Conditioning, Heating, and Refrigeration Institute, Arlington, VA.

ASTM E741-11 (2011). Standard Test Method for Determining Air Change in a Single Zone by Means of a Tracer Gas Dilution, ASTM International, West Conshohocken, PA.

Clarke, J., Strachan, P. A., and Pernot, C. (1993). An approach to the calibration of building energy simulation models. *Transitions-American Society of Heating Refrigerating and Air Conditioning Engineers*, 917–930.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Crawley, D. B., Pedersen, C. O., Lawrie, L. K., and Winkelmann, F. C. (2000). EnergyPlus: Energy simulation program. *ASHRAE Journal*, 42(4).

Griffith, B., N. Long, P. Torcellini, and R. Judkoff, (2008), Methodology for Modeling Building Energy Performance across the Commercial Sector, Technical Report NREL/TP-550-41956, National Renewable Energy Laboratory.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. https://doi.org/dx.doi.org/10.1037/h0071325

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

Jensen, S.O. (1993). Model Validation and Development, In Research Final Report, *PASSYS*, Part II, Commission of the European Communities DGXII.

Judkoff, R., and Neymark, J. (2006). Model validation and testing: the methodological foundation of ASHRAE Standard 140. In *the ASHRAE 2006 Annual Meeting*. Quebec City, Canada.

Judkoff, R., Polly, B., Bianchi, M., Neymark, J. (2010). Building Energy Simulation Test for Existing Homes (BESTEST-EX); Phase 1 Test Procedures: Building Thermal Fabric Cases. Golden, CO: National Renewable Energy Laboratory. NREL/TP-550-47427.

Lomas, K. J., Eppel, H., Martin, C. J., and Bloomfield, D. P. (1997). Empirical validation of building energy simulation programs. *Energy and Buildings*, 26(3), 253–275. https://doi.org/10.1016/S0378-7788(97)00007-8

Macdonald, I. (2002). Quantifying the effects of uncertainty in building simulation. University of Strathclyde. Retrieved from http://www.strath.ac.uk/media/departments/mechanicalengineering/esru/research/phdmphilprojects/macdonald_thesis.pdf

Menberg, K., Heo, Y., and Choudhary, R. (2016). Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. *Energy and Buildings*, 133, 433–445. https://doi.org/10.1016/j.enbuild.2016.10.005

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2), 161–174. https://doi.org/10.2307/1269043

Palomo, E., Marco, J., and Madsem, H. (1991). Method to compare measurements and simulations. In *Proceedings of the 5th International Conference of the International Building Performance Simulation Association*.

Palomo del Barrio, E., and Guyon, G. (2003). Theoretical basis for empirical model validation using parameters space analysis tools. *Energy and Buildings*, 35(10), 985–996. https://doi.org/10.1016/S0378-7788(03)00038-0

Palomo del Barrio, E., and Guyon, G. (2004). Application of parameters space analysis tools for empirical model validation. *Energy and Buildings*, 36(1), 23–33. https://doi.org/10.1016/S0378-7788(03)00039-2

Strachan, P., Monari, F., Kersken, M., and Heusler, I. (2015). IEA annex 58: Full-scale empirical validation of detailed thermal simulation programs. *Energy Procedia*, 78, 3288–3293. https://doi.org/10.1016/j.egypro.2015.11.729

Strachan, P., Svehla, K., Heusler, I., and Kersken, M. (2016). Whole model empirical validation on a full-scale building. *Journal of Building Performance Simulation*, 9(4), 331–350. https://doi.org/10.1080/19401493.2015.1064480

Sun, Y. (2014). Closing the building energy performance gap by improving our predictions. Georgia Institute of Technology.

Walton, G. N. (1981). Passive Solar Extension of the Building Loads Analysis and System Thermodynamics (BLAST) Program, Technical Report, United States Army Construction Engineering Research Laboratory, Champaign, IL.

Wang, Q. (2016). Accuracy, validity and relevance of probabilistic building energy models. Georgia Institute of Technology.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*. Series B: Statistical Methodology, 68(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x