MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

A Virtual Testbed for Robust and Reproducible Calibration of Building Energy Simulation Models

Zhan, Sicheng; Chakrabarty, Ankush; Laughman, Christopher R.; Chong, Adrian

TR2023-114 September 13, 2023

Abstract

A reliable building energy simulation (BES) model is critical for improving building energy performance. While many auto-calibration approaches have been proposed, robust and reproducible BES model calibration remains a challenge due to the lack of a uni- versal evaluation approach and benchmarking frame- work. Therefore, we established a virtual test bed based on DOE prototype buildings to systematically evaluate the calibration results. The Modelica-based testbed enables customized dataset generation and provides the model discrepancy between the calibrated models and the calibration target, which is the key to emulating realistic calibration tasks. We identify three categories of typical pitfalls in BES model calibration and demonstrate them using the virtual testbed. Lastly, a hierarchical model evaluation framework is designed using the testbed for further calibration studies. This study investigates model calibration for buildings from a new perspective and facilitates further research with a standard- ized framework.

IBPSA Building Simulation Conference 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

A Virtual Testbed for Robust and Reproducible Calibration of Building Energy Simulation Models

Sicheng Zhan[1], Ankush Chakrabarty[2], Christopher Laughman[2], Adrian Chong[1]
[1]Department of the Built Environment, National University of Singapore, Singapore
[2]Mitsubishi Electric Research Laboratories, Cambridge, MA, United States

Abstract

A reliable building energy simulation (BES) model is critical for improving building energy performance. While many auto-calibration approaches have been proposed, robust and reproducible BES model calibration remains a challenge due to the lack of a universal evaluation approach and benchmarking framework. Therefore, we established a virtual test bed based on DOE prototype buildings to systematically evaluate the calibration results. The Modelica-based testbed enables customized dataset generation and provides the model discrepancy between the calibrated models and the calibration target, which is the key to emulating realistic calibration tasks. We identify three categories of typical pitfalls in BES model calibration and demonstrate them using the virtual testbed. Lastly, a hierarchical model evaluation framework is designed using the testbed for further calibration studies. This study investigates model calibration for buildings from a new perspective and facilitates further research with a standardized framework.

Highlights

- A high-fidelity Modelica-based virtual testbed is built according to DOE commercial and residential prototype buildings.
- Typical pitfalls in BES model calibration are demonstrated using the virtual testbed.
- A systematic model evaluation framework is designed for robust and reproducible calibration.

Introduction

Improving building energy performance is crucial to reducing carbon emissions and alleviating climate change. Building energy simulation (BES) models are critical in many applications throughout building life cycles, such as retrofit analysis and optimal operations (Hashempour et al., 2020). One prerequisite for making informed decisions in such model-based applications is a reliable model that can represent the buildings in virtual experiments. Although many efforts have been devoted to reducing the discrepancy between the actual and predicted building energy performance, large-scale applications are still hindered by the difficulty of constructing a representative model (Zhan and Chong, 2021).

Calibration of BES models

The difficulty of building energy simulation can be attributed to complicated system dynamics and heterogeneous building characteristics. Correspondingly, the models typically consist of a set of equations that simplify the physics to a certain degree, with a number of parameters that describe the uniqueness of each building. Model calibration is the process of characterizing these unknown parameters to minimize the discrepancy between the prediction and reality (Coakley et al., 2014).

Traditionally, BES models are manually calibrated based on general assumptions and static metadata, which is time-consuming and requires extensive domain knowledge. As more smart meters and IoT sensors are deployed in actual buildings, model calibration can be facilitated by integrating measured data (Chong et al., 2017). Many automatic calibration methods have been proposed during the past decade, including, but not limited to, Bayesian calibration (Chong and Menberg, 2018) and optimization-based methods (Chakrabarty et al., 2021).

Although the effectiveness of these methods has been demonstrated in specific proof-of-concept settings, most existing studies are difficult to reproduce (Chong et al., 2021). This is because the success of a calibration task is subject to many factors. In addition to its own characteristics and disturbances, each building is also distinct regarding the availability and quality of information. These issues need to be comprehensively considered by an expert when configuring the base model and selecting the variables for calibration (Hou et al., 2021). Therefore, although the algorithms automatically find the optimal parameters, the entire calibration procedure is not really automated or scalable. Consequently, model development and calibration remain labor-intensive in practice, which is the most challenging part of model-based applications (Henze, 2013; Blum et al., 2022).

Evaluation of calibration results

The first step to improving the reproducibility of calibration studies is to design a standardized testing framework so that unprejudiced comparisons between different studies can be realized. For example, open-sourced benchmarking datasets have been essential to the thriving artificial intelligence (AI) technologies (Han et al., 2017). In many AI-related fields, such as natural language processing and image recognition, individual samples are self-contained and can be governed by the same distribution. Therefore, data from different sources can be aggregated to form a benchmarking dataset. In contrast, data from different buildings typically observes significant divergence and does not contain all the disturbances (Tian et al., 2018). Therefore, data from dissimilar sources (buildings) cannot be simply put together, making it difficult to establish a comprehensive and robust testing dataset. Considering the cost and difficulty of data acquisition from actual buildings, it is common to utilize synthetic datasets for model training and evaluation.

There are three general approaches to generating synthetic data for buildings: physics-based, data-driven, and hybrid. Klemenjak et al. (2020) created 180 days of power data for two residential households by simulating the usage of 21 electrical appliances; Zhang et al. (2018) applied a Generative Adversarial Network to augment a dataset of time-series total building energy consumption; Roth et al. (2020) produced high-resolution load profiles for every building in a city by integrating annual energy data from actual buildings and hourly energy simulations. Most existing studies of data generation only focused on the electrical meter data, whereas a lot more information (other time-series variable and building metadata) need to be involved when calibrating BES models. The closest synthetic dataset was generated using EnergyPlus[1], where realistic internal and external disturbances were injected to better represent real building operations (Li et al., 2021). However, it only covers one building type at three locations, and HVAC (Heating, Ventilation, and Air Conditioning) system dynamics are unrealistically simplified by EnergyPlus (Wetter et al., 2014). Besides, the dataset was not designed to validate calibrated physical models, which should be done across various operating conditions (Newman et al., 2017). Hence, new synthetic datasets should be dedicated to evaluating the calibration of BES models.

Another pillar of reproducible model calibration is a universal model evaluation approach. A commonly adopted practice is to satisfy the predictive accuracy requirements specified by AHSRAE guideline 14 (ASHRAE, 2005) and IPMVP (IPMVP, 2002). However, the error metrics were found to be frequently miscalculated without a standardized evaluation pipeline (Ramos Ruiz and Fernandez Bandera, 2017). Moreover, it was pointed out that these error metrics were not always reliable. A lower error does not necessarily mean better parameter estimation (ORNL, 2016), and the current threshold of CV(RMSE) and NMBE are insufficient to guarantee the calibration performance (Martínez et al., 2020). It has also been recognized that modeling and calibration should be streamlined to the application being studied (Trčka and Hensen, 2010). Therefore, the focus of BES calibration should shift from reducing predictive errors for a single building to developing robust and generalizable calibration strategies.

Objectives and synopsis

The objective of this study is to address the aforementioned issues in BES model calibration by establishing a standardized virtual platform. A robust and reproducible calibration framework should be able to produce reliable BES models without the supplementary intervention of domain experts. Therefore, the calibration performance should be comprehensively examined with respect to various operating conditions and heterogeneous buildings. The synthetic dataset, generated by these higher-fidelity emulators, is the first dataset dedicated to BES model calibration.

In the rest of this paper, we first formalize the problem of BES model calibration and pinpoint the affecting factors. Accordingly, the virtual testbed and the synthetic dataset are introduced, accompanied by some descriptive simulation results. The testbed is then used to demonstrate several typical pitfalls in the current practice of model calibration. Lastly, a hierarchical evaluation system is defined for BES model calibration.

Problem statement

The problem of calibrating BES models can be formulated as Equation 1:

$$\begin{aligned} \theta^* &= arg\min_{\theta\in\Theta}(J(y, y^*)) \\ s.t. \quad y &= \mathcal{M}(x, u, \theta) \end{aligned} \tag{1}$$

The true parameters θ^* are sought by minimizing the objective function J. The first element y is the target outputs of BES model $\mathcal{M}$, where θ is the model parameters to be calibrated, x is the model inputs such as weather conditions, and u is the control variables such as setpoints. The second element y^* is the ground truth of y, and the calibration is subject to the admissible parameter range Θ. All of these components have a profound impact on the calibration results and need to be carefully configured when defining the problem. We summarize these affecting factors and related potential issues from four aspects:

- θ and Θ: Given a BES model, a necessary step

[1] https://energyplus.net/

is to decide which parameters to calibrate and the corresponding range. For Bayesian calibration, Θ is given in the form of prior distributions. BES models usually have a large number of parameters, and it is impossible to calibrate all of them. Calibrating more parameters or at a higher resolution does not always yield better results (Chong et al., 2021). Furthermore, the selection of parameters should be tied to factors such as building characteristics, application purposes, and data availability. Therefore, systematic parameter selection should be part of a robust calibration procedure. Sensitivity analysis is a well-established approach to selecting the parameters (Tian, 2013). However, due to the absence of a generalizable guideline, many past studies did not perform a rigorous sensitivity analysis.

- J and y: The objective functions quantify the distance between y and y^*, sometimes transformed to be compatible with the optimization or calibration algorithms. Theoretically, the model outputs of interest should be specified based on the calibrated parameters and the downstream applications. For example, calibrating too many parameters with limited outputs could lead to identifiability issues, and the spatial and temporal resolution of calibration should be the same as the application scenarios. However, the choice of y is often made at the beginning of a project, dominated by the availability of data. Over 90% of calibration studies used less than three outputs (Chong et al., 2021).
- y^* and $\mathcal{M}$: When calibrating BES models for actual buildings, ground truth y^* is measured data, and the model discrepancy is an impactful factor that is difficult to account for. Building dynamics are complicated, and it is impossible to fully monitor the disturbances. Therefore, BES models can only represent the primary behavior and are invariably simplified. For example, room temperature is modeled as a node of well-mixed air, and many parameters, such as power densities, are often defined universally across the entire building. In such cases, θ^* sometimes have abstract values, and $\mathcal{M}$ typically cannot perfectly match y^* (Zhan et al., 2022).

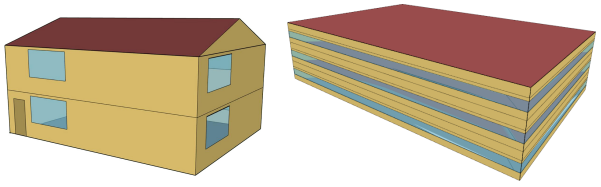
  Considering the flexibility and lower cost of simulations, many studies used synthetic instead of measured data. As part of the synthetic dataset, the ground truth $y^* = \mathcal{M}'(x, u, \theta^*)$, where $\mathcal{M}'$ is the emulator that generated the training data. If the model to be calibrated $\mathcal{M}$ has the same form as $\mathcal{M}'$, such as EnergyPlus, it becomes an idealized case that is not affected by the model discrepancy. In such cases, the calibration problem can be well-constrained so that θ^* can be exactly identified ($\mathcal{M} = \mathcal{M}'$), which is unlikely to happen in reality. To avoid this problem and mimic the potential model discrepancy, the emulator $\mathcal{M}'$ should have higher fidelity than $\mathcal{M}$.
- x and u: Most existing datasets are collected from a period of normal building operations when the x and u are restricted within a relatively narrow range. In contrast, many downstream applications require the calibrated model to extrapolate beyond normal operating conditions, such as predicting energy performance under climate change scenarios and exploring some new control actions. To ensure reliability, the calibrated BES models should be comprehensively evaluated by varying x and u to cover these potential prediction cases.

Design of the virtual testbed

Based on the previous discussion, the proposed virtual testbed is designed to have three key features:

1. Validated typical building models, covering a certain level of inter-building heterogeneity.
2. Pairs of emulators and candidate models (for calibration) to account for the model discrepancy.
3. Flexible simulation interfaces to customize the tests for potential applications.

The U.S. Department of Energy (DOE) provides a series of prototype models for commercial and residential building types[2]. We select the most popular building type from each category, medium office and single-family house, to form the virtual testbed. Each building type includes a group of models that serve the first key feature. The models in each group have a standard geometry and zoning (Figure 1), with various constructions for 15 different climate zones based on the 2018 International Energy Conservation Code[2]. The variability mainly lies in the boundary conditions, the thermal properties of constructions, and the HVAC system specifications.



(a) Single-family house. *(b) Medium office.*

Figure 1: Geometry of the building models rendered by SketchUp.

Table 1 summarize the basic information of these two models. The single-family house has two floors but is modeled as one thermal zone, with an unconditioned attic. The living unit is conditioned by a fan coil unit, with a single-speed fan, a direct expansion cooling coil, and an electric heating coil. The medium office is

[2]https://www.energycodes.gov/prototype-building-models

a larger-scale multi-zone building. Each floor has four perimeter zones, one core zone, and an unconditioned plenum. Each floor is conditioned by one AHU, with a direct expansion cooling coil, a natural gas heating coil, and variable air volume reheat terminals. More details about the prototype models can be found in the original DOE models[2].

Table 1: Basic information of the building models.

	Single house	**Medium office**
No. levels	2	3
No. conditioned zones	1	15
Total floor area (m^2)	110	4988
No. unconditioned zones	1	3
Supply air fan	On/off	Variable speed
Cooling coil	Direct expansion	
Main heating coil	Electric	Natural gas
Reheat coil	None	Electric

Modelica-based emulators

The original DOE models are given in EnergyPlus. We convert the models into a Modelica library to enable the second and the third key features. Compared with EnergyPlus, Modelica captures the building side dynamics in a similar way but provides the capability of modeling higher-fidelity HVAC systems. The Modelica models were built using the same metadata as in the original models. Figure 2 presents a comparison of free-floating temperature predictions, where a pair of Modelica and EnergyPlus models are well-aligned given the same metadata and configurations. For HVAC systems, the system performance is captured in a transient, instead of steady, state, and the control strategies are incorporated.

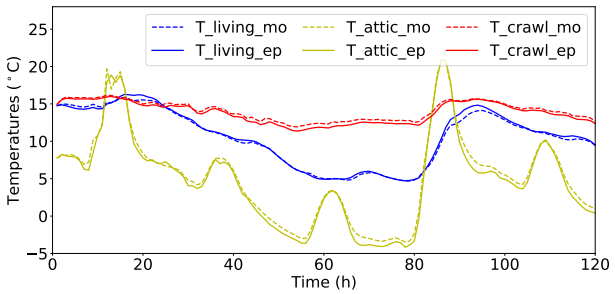


Figure 2: Free-floating temperature predictions (`living` unit, `attic`, and `crawlspace`) of a pair of Modelica (`mo`) and EnergyPlus (`ep`) models.

The Modelica library `DOE_testbed` is constructed in an object-oriented approach, as shown in Figure 3. The `BaseClasses` are created respectively for `SingleHouse` and `MultiFamily`, and the building models are constructed for each climate zone by switching the course data classes. The library is built upon Modelica Standard Library 3.2.3 and Buildings 7.0.0[3], compiled into fmu[4] using Pymodelica[5], and simulated with FMPy[6].

The high-fidelity Modelica models serve as the emulators to generate synthetic data, and the corresponding EnergyPlus models (assumed to have unknown parameters), or other models of lower fidelity, are to be calibrated. Thereby, the model discrepancy can be introduced to make a more realistic calibration problem. Besides, the fmu-based toolchain can be implemented in a standard way and is compatible with other calibration tools. It is also convenient to examine if a calibrated model will suffice for potential applications. For example, the control actions optimized by the calibrated model can be applied back to the emulator for evaluation.

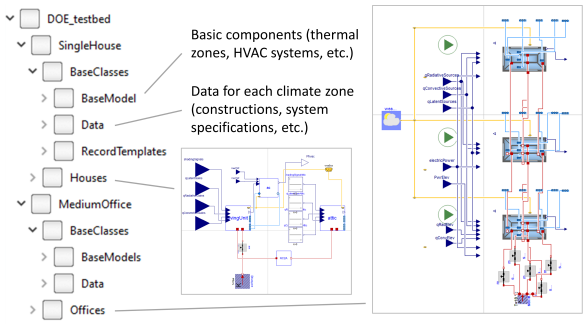


Figure 3: Class hierarchy of the Modelica library.

Figure 4 uses the single-family houses to show the significant variability of the emulators in terms of the building characteristics and the simulation results. Because of the difference in outdoor conditions and building thermal properties, the free-floating living unit temperature in January varies from -10 to over 30°C. When the air conditioning is activated with 22°C heating setpoint and 24°C cooling setpoint, the energy consumption correspondingly changes across climate zones. When it comes to calibration, this variation could seriously affect the parameter selection and the results.

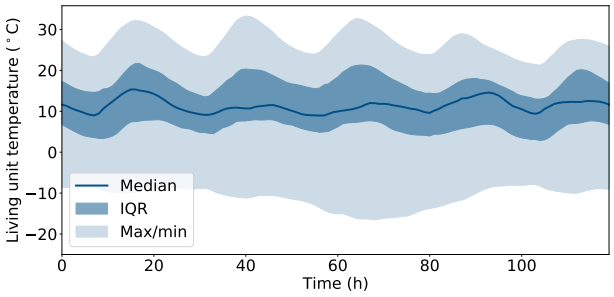
Pitfalls in calibration

Misuse of error metrics

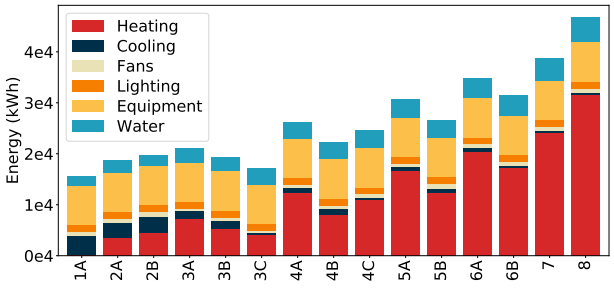
In the literature about BES model calibration, the model evaluation approach that is usually adopted is to compute the predictive errors and compare them with the standard thresholds. For example, IPMVP requires the CV(RMSE) of monthly utility bills to be lower than 5% and that of hourly energy use to be lower than 20% (IPMVP, 2002). Models that fulfill these standards are usually considered "calibrated", which could easily be unreliable.

We manifest this problem with a simple calibration task using the virtual testbed. As sub-metering is not

[3]https://github.com/lbl-srg/modelica-buildings
[4]https://fmi-standard.org/
[5]https://pypi.org/project/PyModelica/
[6]https://github.com/CATIA-Systems/FMPy



(a) Free-floating living unit temperature.



(b) Disaggregated annual energy consumption.

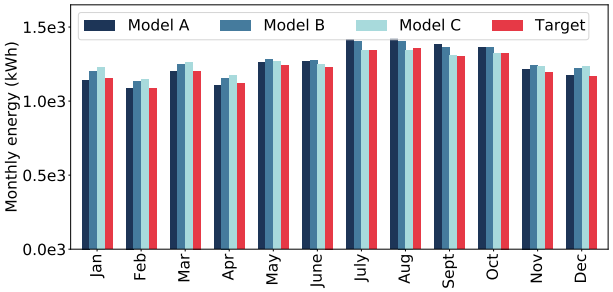
Figure 4: Emulation results of the single-family houses in 15 climate zones.

available in most buildings, it is common to calibrate several parameters only based on the total building energy consumption. Accordingly, EnergyPlus models A, B, and C were calibrated based on the total energy consumption throughout the year. The models represent a single-family house in Hawaii, and only two parameters were calibrated: electric power density and nominal cooling coil efficiency. As shown in Table 2, all three models had similar monthly and hourly total energy consumption errors, lower than the threshold. However, a large variability can be observed in the predicted cooling consumption, the hourly CV(RMSE) of which went up to 77.19%.

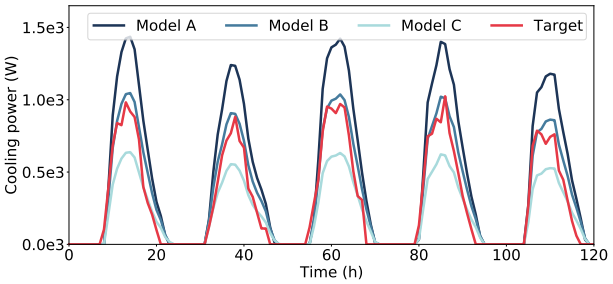
Table 2: Predictive errors of three calibrated models.

CV(RMSE)	A	B	C
Monthly total	3.36%	3.96%	3.69%
Hourly total	19.98%	11.82%	16.16%
Hourly cooling	77.19%	27.64%	51.70%

Figure 5 plots the prediction results with the ground truth. While the monthly data generally followed the trend, the instant cooling power deviated by as large as 50%. Essentially, the total energy consumption is composed of various end uses with different ratios, and the error in one category could be compensated for by other categories. As stated in the documents, the error thresholds are meant for forward calibration problems, where the parameters are mostly justified by metadata or experiments. For a reverse problem that is based on time-series measured data, simply fulfilling the error requirements could be misleading. Such an error in end-use prediction could harm further applications such as assessing energy conservation measures (ECMs).



(a) Monthly total energy consumption.



(b) Hourly cooling power.

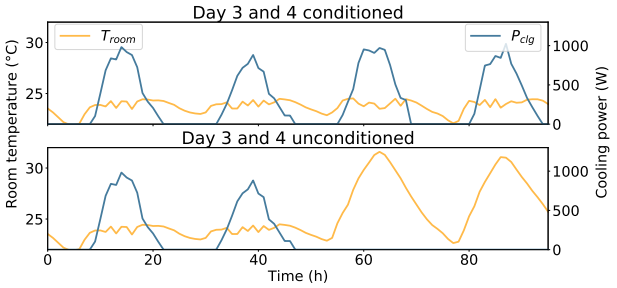
Figure 5: Prediction results of three calibrated models.

Identifiability issues

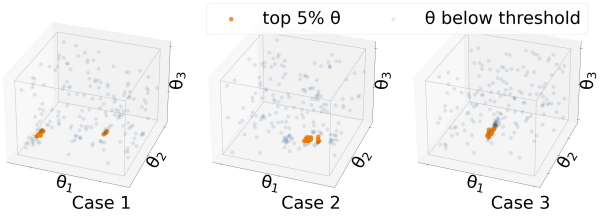
Another problem that commonly happens in calibration studies is the identifiability issue, which is also the underlying cause of the previous example. When only the total energy is provided, the data is not informative enough to estimate all the parameters. To separate cooling from total energy consumption is straightforward, taking just one additional electricity meter. Yet, a realistic calibration task would have more unknown parameters such as the envelope properties and the operating schedules. In those cases, more information, other than the cooling power data, would be needed to improve the identifiability.

We conducted a series of optimization-based optimization experiments to manifest this problem. Three parameters related to the building cooling performance were selected, including the solar heat gain coefficient of windows, the overall effective leakage area, and the cooling coil efficiency. The calibration was realized by Bayesian Optimization, consisting of 150 exploratory simulations with random parameters and 50 iterations to minimize the mean squared error. Figure 6 visualizes the four days of hourly synthetic data under two operation schemes and the parameter space of the three cases, where the optimal parameter sets are highlighted in orange.

The simplest case 1 only used the cooling power with the air conditioning system always available, yielding two clusters of parameters that cannot be distinguished by the predictive error. Case 2 still used the data when the house is constantly conditioned but



(a) Results of two data generation schemes.



(b) Parameter space of three cases.

Figure 6: Synthetic data and calibration results of three cases with increasing identifiability.

appended the room temperature as another output. The additional information improved the identifiability, reflected by the closer but still separated clusters. Ultimately, Case 3 generated a new dataset by disabling the cooling on the third and fourth days. Using the cooling power and room temperature of this dataset, a unique parameter set was identified.

This example indicates the importance of carefully designing the calibration task, especially when an optimization-based approach is adopted. Calibrating against multiple outputs introduces more information about building physics and is usually helpful. Nonetheless, the complicated mutual effect between parameters could still be tricky to resolve. Martínez et al. (2020) also reported that adding temperature data as output did not help the calibration. In such cases, a potential solution is to include more variability when generating the data.

Model discrepancy

As important as it is to improve identifiability, an identifiable model does not guarantee the "correctness" of calibrated parameters. In the previous experiment, the three cases resulted in very disparate parameter combinations. Although getting closer, even case 3 did not identify the same parameter as the original specifications. This is attributed to the model discrepancy, which is inevitable in reality. As the physical processes described by the model are different from the target (emulator or real building), the original parameters will produce a non-zero predictive error that can potentially be reduced by altering the parameters. There are many sources of model discrepancies, and several are spotted in the virtual testbed of single-family houses:

- EnergyPlus models an idealized steady-state HVAC system, and the room temperature will be constantly at the setpoint as long as the cooling is activated. In contrast, a realistic HVAC system dynamically tracks the setpoints, and the temperature will slightly fluctuate around the setpoint as shown in Figure 6.a. Therefore, EnergyPlus models can never perfectly predict the temperature trend.

- The heat convection of exterior building surfaces in EnergyPlus is governed by the object `SurfaceConvectionAlgorithm:Outside`, which is set as `DOE-2` by default in the DOE prototype models. `DOE-2` is a simplified algorithm that estimates the convective heat transfer coefficient based on surface roughness and local wind speed. The prediction can be improved simply by changing this object to `TARP`, which is a more complicated and robust algorithm.

- There are many electrical appliances in residential buildings, with different schedules, heat loss ratios, and energy intensities. The emulators incorporate these objects according to the prototype models. However, it is impractical to gather all the information in reality. While the electric power can be monitored with smart meters, it is impossible to precisely capture the related internal heat gains. An acceptable compromise is to approximate the effect with a lumped equipment object based on the standards.

On the other hand, not having physically meaningful parameters does not mean the calibrated model cannot be used. In fact, calibrating more parameters provides higher degrees of freedom to minimize the predictive error. While it is necessary to carefully examine the models to avoid potential overfitting problems, a model with "wrong" parameters may outperform one with physically meaningful parameters. For example, the three calibrated parameters of case 3 in Figure 6 are 10-20% off the original values. Yet, although the calibration only used four days of data, the annual prediction errors are lower than the original model, with a monthly CV(RMSE) of 1.54% and an hourly CV(RMSE) of 11.77%.

Note that the actual model discrepancy in practice is usually much larger than the toy test case here, and the amount of reliable metadata is much less. Models of higher fidelity are more demanding about the historical data needed to form a proper calibration problem. Meanwhile, when the model fidelity is too low to capture the dominant physical processes, any calibration will not be able to close the gap. Hence, it is essential to carefully determine the model fidelity considering the data availability.

Hierarchical model evaluation

Based on the virtual testbed and the previous discussion, three levels of model evaluation are designed for future BES calibration studies. As the ground truth of parameters is practically unknown, the evaluation is realized by prediction tests. Figure 7 illustrates the workflow of a calibration study using the virtual testbed, with the schematics of the hierarchical tests. It starts with configuring the calibration task and properly documenting the design. (Chong et al., 2021) proposed a documentation checklist to promote reproducibility, including building information, input and output sources, calibration parameters, modeling assumptions, etc.

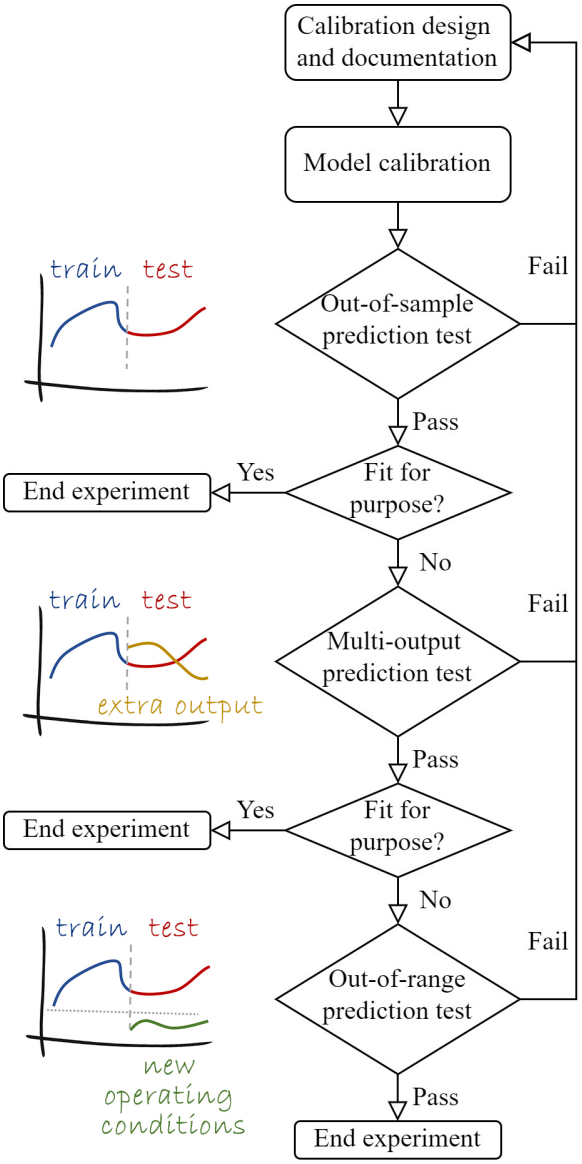


Figure 7: Hierarchical evaluation workflow using the virtual testbed.

After calibration, the first level of evaluation that every study should conduct is the out-of-sample prediction tests. Although separating the dataset into training and testing has been almost mandatory for machine learning studies, past BES calibration studies commonly used just one dataset, probably carried over from the forward calibration regime. There are two dimensions of out-of-sample tests. Temporally, the calibrated model should be tested during a period that is outside of the training dataset and covers different seasons, especially if the training is done with a few days of data. Spatially, the calibration procedure should be repeated at a new location, as the resulting performance could vary under different conditions. With the virtual testbed, this new location will be selected by adding five to (or subtracting from) the original climate zone code.

While three levels of tests are all included in the workflow, the experiment can stop after the first level if the calibrated model already satisfies the application requirements. The multi-output prediction test is to be done only if required by the downstream application. Apart from the most interested energy consumption and room temperature, applications such as optimal control may involve other variables such as HVAC operating conditions or non-shiftable electricity usage. These scenarios ask for a more precise representation of the physical process, which needs to be explicitly tested. Similarly, whether and how to conduct the out-of-range test is also subject to modeling purposes. The potential scenarios, such as climate change or new operating conditions, can be emulated on the virtual testbed to generate a fit-for-purpose testing dataset. Thereby, the calibration results will be reliable for the intended purpose.

Conclusion

This study aims to design a benchmarking framework to improve the robustness and reproducibility of model calibration for buildings. First, we discuss the affecting factors of a calibration task and specify the key feature of such a testing platform. Accordingly, the Modelica-based virtual testbed is built according to the DOE prototype buildings. Using the testbed, three pitfalls in model calibration are demonstrated, and a standardized calibration and evaluation workflow is designed for future studies.

Acknowledgment

This research project is supported by the National Research Foundation, and the Ministry of National Development, Singapore, under its Cities of Tomorrow R&D Programme (CoT Award COT-V4-2020-5) and Johnson Controls (No. A-8001200-00-00).

References

American Society of Heating, Refrigerating and Air-Conditioning Engineers (2005). *ASHRAE's Guideline 14-2002 for Measurement of Energy and Demand Savings: How to Determine what was really saved by the retrofit*.

Blum, D., Z. Wang, C. Weyandt, D. Kim, M. Wetter, T. Hong, and M. A. Piette (2022). Field demonstration and implementation analysis of model predictive control in an office hvac system. *Applied Energy 318*, 119104.

Chakrabarty, A., E. Maddalena, H. Qiao, and C. Laughman (2021). Scalable bayesian optimization for model calibration: Case study on coupled building and hvac dynamics. *Energy and Buildings 253*, 111460.

Chong, A., G. Augenbroe, and D. Yan (2021). Occupancy data at different spatial resolutions: Building energy performance and model calibration. *Applied Energy 286*, 116492.

Chong, A., Y. Gu, and H. Jia (2021). Calibrating building energy simulation models: A review of the basics to guide future work. *Energy and Buildings 253*, 111533.

Chong, A., K. P. Lam, M. Pozzi, and J. Yang (2017). Bayesian calibration of building energy models with large datasets. *Energy and Buildings 154*, 343–355.

Chong, A. and K. Menberg (2018). Guidelines for the bayesian calibration of building energy models. *Energy and Buildings 174*, 527–547.

Coakley, D., P. Raftery, and M. Keane (2014). A review of methods to match building energy simulation models to measured data. *Renewable and sustainable energy reviews 37*, 123–141.

Han, R., L. K. John, and J. Zhan (2017). Benchmarking big data systems: A review. *IEEE Transactions on Services Computing 11*(3), 580–597.

Hashempour, N., R. Taherkhani, and M. Mahdikhani (2020). Energy performance optimization of existing buildings: A literature review. *Sustainable Cities and Society 54*, 101967.

Henze, G. P. (2013). Model predictive control for buildings: a quantum leap? *Journal of Building Performance Simulation 6*(3), 157–158.

Hou, D., I. Hassan, and L. Wang (2021). Review on building energy model calibration by bayesian inference. *Renewable and Sustainable Energy Reviews 143*, 110930.

International Performance Measurement & Verification Protocol (2002). *International performance measurement and verification protocol: Concepts and Options for Determining Energy and Water Savings-Vol. I*.

Klemenjak, C., C. Kovatsch, M. Herold, and W. Elmenreich (2020). A synthetic energy dataset for non-intrusive load monitoring in households. *Scientific data 7*(1), 108.

Li, H., Z. Wang, and T. Hong (2021). A synthetic building operation dataset. *Scientific data 8*(1), 213.

Martínez, S., P. Eguía, E. Granada, A. Moazami, and M. Hamdy (2020). A performance comparison of multi-objective optimization-based approaches for calibrating white-box building energy models. *Energy and Buildings 216*, 109942.

Newman, A. J., N. Mizukami, M. P. Clark, A. W. Wood, B. Nijssen, and G. Nearing (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology 18*(8), 2215–2225.

Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States) (2016). *Suitability of ASHRAE guideline 14 metrics for calibration*.

Ramos Ruiz, G. and C. Fernandez Bandera (2017). Validation of calibrated energy models: Common errors. *Energies 10*(10), 1587.

Roth, J., A. Martin, C. Miller, and R. K. Jain (2020). Syncity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods. *Applied Energy 280*, 115981.

Tian, W. (2013). A review of sensitivity analysis methods in building energy analysis. *Renewable and sustainable energy reviews 20*, 411–419.

Tian, W., Y. Heo, P. De Wilde, Z. Li, D. Yan, C. S. Park, X. Feng, and G. Augenbroe (2018). A review of uncertainty analysis in building energy assessment. *Renewable and Sustainable Energy Reviews 93*, 285–301.

Trčka, M. and J. L. Hensen (2010). Overview of hvac system simulation. *Automation in construction 19*(2), 93–99.

Wetter, M., W. Zuo, T. S. Nouidui, and X. Pang (2014). Modelica buildings library. *Journal of Building Performance Simulation 7*(4), 253–270.

Zhan, S. and A. Chong (2021). Data requirements and performance evaluation of model predictive control in buildings: A modeling perspective. *Renewable and Sustainable Energy Reviews 142*, 110835.

Zhan, S., Y. Lei, Y. Jin, D. Yan, and A. Chong (2022). Impact of occupant related data on identification and model predictive control for buildings. *Applied Energy 323*, 119580.

Zhang, C., S. R. Kuppannagari, R. Kannan, and V. K. Prasanna (2018). Generative adversarial network for synthetic time series data generation in smart grids. In *2018 SmartGridComm*, pp. 1–6. IEEE.