Calibration of Building Energy Computer Models via Bias-Corrected Iteratively Reweighted Least Squares Method

Cheoljoon Jeong[a], Eunshin Byon[a]

[a] *University of Michigan, 1205 Beal Avenue, Ann Arbor, 48105, MI, USA*

Abstract

As the building sector contributes approximately three-quarters of the U.S. electricity load, analyzing buildings' energy consumption patterns and establishing their effective operational strategy become of great importance. To achieve those goals, a physics-based building energy model (BEM), which can simulate a building's energy demand under various weather conditions and operational scenarios, has been developed. To obtain accurate simulation outputs, it is necessary to calibrate some parameters required for the BEM's pre-configuration. The BEM calibration is usually accomplished by matching the simulated energy use with the measured one. However, even with the efforts to calibrate the BEM at best, a systematic discrepancy between the two quantities is often observed, preventing the precise estimation of the energy demand. Such discrepancy is referred to as bias in this study. We present a new calibration approach that models the discrepancy to correct the relationship between the simulated and measured energy use. We show that our bias correction can improve predictive performance. Additionally, we observe the heterogeneous variance in the electricity loads, especially in the afternoon hours, which often reduces prediction accuracy and increases uncertainty. To address this issue, we incorporate heterogeneous weights into the least squares loss function. To implement the bias-correction procedure with the weighted least squares formulation, we propose a newly devised iteratively reweighted least squares algorithm. The effectiveness of the proposed calibration methodology is evaluated with a real-world dataset collected from a residential building in Texas.

Keywords: Bias Correction, Building Energy Simulation, Heteroskedasticity, Weighted Least Squares

1. Introduction

As the building sector accounts for approximately 40% of the primary energy use and 72% of electricity loads in the U.S. (U.S. Energy Information Administration 2015), it becomes vital to analyze the energy consumption patterns of buildings and improve their energy efficiency. In response, physics-based building energy models (BEMs) have been developed, which are capable of simulating a building's energy consumption, including electricity and gas or steam energy, by maneuvering various conditions of heating, ventilation, and cooling (HVAC), lighting, and plug and process loads under various weather conditions, operational schedules, and building geometry. Another usage of the BEMs includes, but not limited to, HVAC system design and operation, retrofit analysis, and architectural design. Among several BEMs (or building energy simulation engines), the U.S. Department of Energy's National Renewable Energy Laboratories developed a simulator, called EnergyPlus (U.S. Department of Energy 2019), which has gained much popularity in evaluating a building's energy performance in the literature (Chong et al. 2021).

However, several studies report a considerable discrepancy between simulated and actual energy use, raising concerns about the model's reliability in the building sector (Turner et al. 2008, Mantesi et al. 2018). Due to the advances in smart metering and industrial internet of things (IIoT) technologies, this discrepancy becomes more evident (Chong et al. 2021). Several reasons for the discrepancy have been discussed in prior studies (De Wit and Augenbroe 2002, Menezes et al. 2012, Chakrabarty et al. 2021). Assuming that building specifications are sufficiently detailed and data accuracy is guaranteed, that is, the data source and measurement techniques are reliable, the discrepancy could result from (i) parameter uncertainty arising from the fact that the initial (or default) values of simulation parameters in BEM descriptions would not accurately reflect the underlying physics, and (ii) model inadequacy caused by simplifications and abstractions of a real building's energy systems. Thus, the calibration procedure for reducing the discrepancy

between the simulated energy consumption and actual observations, along with suitable uncertainty analysis, is essential to enhancing model reliability. The International Energy Agency's Energy in Buildings and Communities (IEA-EBC) Annex 53 also discussed the importance of model calibration and uncertainty quantification to obtain a credible BEM (Yoshino et al. 2017).

Model calibration is usually accomplished by adjusting simulation parameters so that the simulated values of energy consumption are closely aligned with the actual observations. It is also known as parameter calibration in the literature (Xu et al. 2021, Liu et al. 2021, Jeong et al. 2023), aiming to capture a target building's real physical dynamics. Once calibrated, the parameters are not only useful for the accurate simulation of energy use but also enable us to infer information about the states in the building's energy system and their physical implications.

Even with the efforts to calibrate the BEM parameters as accurately as possible, a systematic discrepancy between the two series of simulated and actual energy consumption often exists. This discrepancy is referred to as bias in this study. The bias may exhibit a distinct pattern, such as a daily cycle (Jang et al. 2023), which needs to be taken into consideration in the calibration procedure. Statistical approaches that account for the bias have been discussed in the Bayesian calibration literature (Kennedy and O'Hagan 2001). In fact, Bayesian calibration has been widely used with its uncertainty quantification capabilities in the building energy literature (Coakley et al. 2014, Chong and Menberg 2018). Despite its popularity, one of its biggest drawbacks is the high computational cost, particularly when dealing with high-resolution data, such as hourly and sub-hourly data. Thus, its application has been limited to low-resolution aggregated data such as weekly (Kristensen et al. 2017), monthly (Heo et al. 2015, Li et al. 2016, Kim and Park 2016, Tian et al. 2016, Sokol et al. 2017), or annual (Booth et al. 2013) data. This might also be a common practice since electricity and gas or steam data were usually obtained from utility providers who typically provided aggregated monthly data (Chong et al. 2021). Unfortunately, the low-resolution data may lose useful information in data aggregation.

To alleviate computational burden, Li et al. (2016) suggested a lightweight Bayesian calibration approach that employs a linear regression emulator. Menberg et al. (2017) applied Hamiltonian Monte Carlo to enhance posterior estimation efficiency. In spite of these advancements, computational demands still remain a hurdle in the Bayesian calibration, limiting its practicality to small-size datasets. For instance, Jeong et al. (2023) reported that it took several days to calibrate multiple BEM parameters with weekly data using the lightweight Bayesian approach and that, despite the long computation time, the calibration results were not informative, presumably due to the information loss during data aggregation.

Recently, advances in smart metering and IIoT technologies enable us to access high-resolution data with high precision. This big data stream brings us the opportunity to implement optimization-based methods for calibration. Jeong et al. (2023) presented a BEM calibration method that utilizes a gradient-based optimization technique with hourly data. Chakrabarty et al. (2021) applied Bayesian optimization (BO) with relatively large-size datasets for the BEM calibration. However, these studies did not consider the bias in their procedure, possibly leading to an incomplete relationship between the simulated and actual energy consumption.

Additionally, we note that the heterogeneous variance in the electricity loads is particularly pronounced in the afternoon in the case study considered in this article. Although there is a general electricity consumption pattern in the afternoon, the specific hour-by-hour electricity loads show very varied patterns each day (see more details in Section 3). Such heteroskedasticity is a violation of the typical constant variance assumption in the literature, which could reduce prediction accuracy and increase estimation uncertainty. However, no studies in the BEM calibration literature account for the heterogeneous variance. To address this issue, we introduce heterogeneous weights within the least squares loss function between the two time series, leading to the weighted least squares formulation.

To address these challenges, this study presents a new calibration approach that models the bias to correct the relationship between the simulated and measured energy use and, at the same time, introduces heterogeneous weights within the loss function between the two time series of simulated and actual ones in order to mitigate heteroskedasticity. To calibrate the BEM parameters and estimate other model parameters in an integrative framework, we propose a newly devised iteratively reweighted least squares (IRLS) algorithm. To the best of our knowledge, this is the first research to present the IRLS method in conjunction with the bias-correction procedure in the BEM calibration.

We summarize our contributions as follows. First, in order to calibrate the BEM parameters and align the BEM outputs with actual electricity use, we provide a new modeling approach that debiases the BEM outputs while accounting for the heterogeneous variance. Second, we present a new procedure to estimate the BEM and other model parameters integratively. Third, utilizing the fact that the calibrated parameter values become maximum likelihood

(ML) estimates, we quantify the estimation uncertainties and construct asymptotically valid confidence intervals for the BEM parameters. Lastly, we conduct a case study using a real-world electricity consumption dataset collected from a residential building in Texas to evaluate the effectiveness of the proposed calibration methodology.

Notably, our case study demonstrates that the proposed method, when compared to other alternatives, significantly improves prediction accuracy for the electricity demands while simultaneously reducing the uncertainties of the calibrated parameters, satisfying the industry guidelines and protocol. Moreover, our uncertainty quantification procedure results in constructing narrower confidence intervals, i.e., smaller uncertainties, compared to alternative methods.

The rest of this paper is organized as follows. Section 2 introduces a linear linkage model that assumes no bias in the model. Section 3 formulates a calibration problem that debiases the BEM outputs, provides an approach to mitigate heteroskedasticity, and designs a new bias-correction algorithm for parameter calibration within the IRLS framework. Section 4 demonstrates the superiority of our proposed approach through the real-world BEM calibration case study for a residential building in Texas. Section 5 provides concluding remarks.

2. A Linear Linkage Model without Bias Assumption

Before discussing our proposed approach, we first present the widely used linear linkage model and its limitations. Let $\mathbf{x}_t \in \mathbb{R}^{M_\mathrm{x}}$ denote a vector of M_x physically observable input variables for the BEM, such as dry-bulb temperature, wind speed, solar radiation, and relative humidity, collected in a building's surrounding area at time t for all $t = 1, \ldots, T$, where T is the number of historical time-series observations. Let $\boldsymbol{\theta} \in \mathbb{R}^{P_\theta}$ denote a vector of calibration parameters where the range of the ith parameter is $[a_i, b_i]$ with a_i and b_i known constants for all $i = 1, \ldots, P_\theta$, implying that the domain of $\boldsymbol{\theta}$ is a hyperrectangle $\boldsymbol{\Theta} := \prod_{i=1}^{P_\theta} [a_i, b_i]$. The parameters we consider in this study are those related to envelop (e.g., solar transmittance), zone (e.g., air flow rate), and HVAC (e.g., cooling coefficient of performance and component capacity), but one can select others according to the calibration goal and scenario. Let $y(\mathbf{x}_t)$ denote a real-valued noisy field observation for an input $\mathbf{x}_t$, such as the measurement of energy consumption, including gas or electricity use, within the building. In this study, we consider the hourly electricity consumption. Let $\eta(\mathbf{x}_t; \boldsymbol{\theta})$ denote a real-valued output from the BEM, such as the energy consumption value simulated by the BEM. We consider the deterministic computer model that generates a fixed output given $\mathbf{x}_t$.

This study aims to align $\eta(\mathbf{x}_t; \boldsymbol{\theta})$ with $y(\mathbf{x}_t)$ reasonably well to accurately predict electricity consumption through the BEM simulation. To do this, the unknown parameters $\boldsymbol{\theta}$ should be properly estimated using operational data. Let us first consider that a simulator precisely represents the underlying physical process when the true (or correct) parameters are used. This type of computer model is called a *perfect* computer model in the literature (Tuo and Wu 2015). To connect field observations with the computer model outputs, a linear linkage model has been proposed in the literature (Higdon et al. 2004) as follows:

$$y(\mathbf{x}_t) = \eta(\mathbf{x}_t; \boldsymbol{\theta}) + \epsilon_t, \quad \forall t = 1, \ldots, T, \tag{1}$$

where an observation error ϵ_t is assumed to be an independent and identically distributed (iid) random variable that follows a normal distribution with mean zero and variance σ^2, concisely, $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Let $\boldsymbol{\theta}^*$ denote the correct parameter values. Mathematically, a perfect computer model implies that $E[y(\mathbf{x}_t)] = \eta(\mathbf{x}_t; \boldsymbol{\theta}^*)$ holds for all t. In other words, there is no bias and the BEM simulator generates unbiased results over time when $\boldsymbol{\theta}$ is correctly estimated. Under this unbiasedness assumption, the calibration problem is to identify the estimator $\hat{\boldsymbol{\theta}}$ that minimizes the difference between $y(\mathbf{x}_t)$ and $\eta(\mathbf{x}_t; \boldsymbol{\theta})$ for all t (Jeong et al. 2023). The difference can be quantified by some difference measures, among which the following mean squared error (MSE) is widely employed.

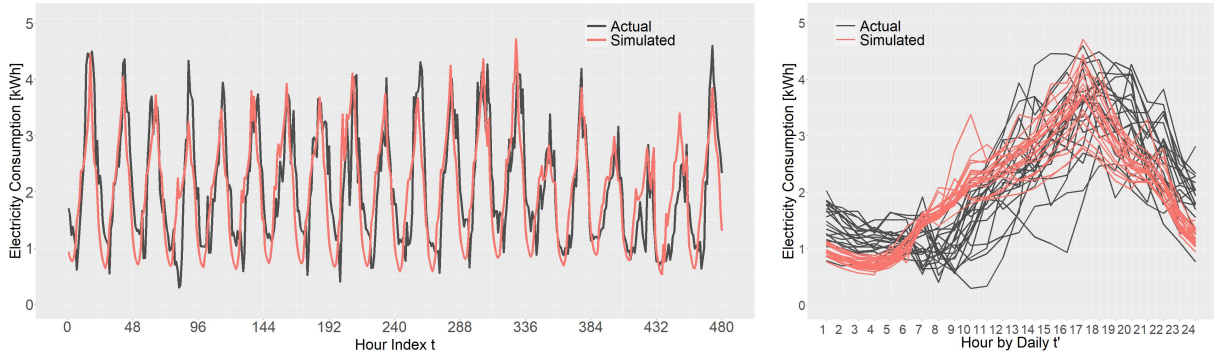
$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{T} \sum_{t=1}^{T} (y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \boldsymbol{\theta}))^2. \tag{2}$$

When a small number of field observations are available and/or the computer model is expensive to run, surrogate-based approaches, such as Bayesian calibration (Higdon et al. 2004) and L_2 calibration (Tuo and Wu 2015), have been used in the literature to accommodate scarce data. They typically pre-specify design points $\mathbf{x}_t$ and $\boldsymbol{\theta}$ and estimate the true process $\zeta(\mathbf{x}_t)$, where $y(\mathbf{x}_t) = \zeta(\mathbf{x}_t) + \epsilon_t$ and/or $\eta(\mathbf{x}_t; \boldsymbol{\theta})$ using those design points. They are particularly useful

when there is a limited amount of available data (Liu et al. 2021). Recently, some limitations of these surrogate-based approaches are discussed when a sufficient number of field observations are available and computer models are relatively cheap to run, and new approaches are proposed using the nonlinear optimization techniques (Liu et al. 2021, Xu et al. 2021, Jeong et al. 2023, Jain et al. 2023). They do not pre-design the input data, but generate them *on the fly* as they learn. Their methods are especially useful when data generation from the computer model is not computationally intensive, e.g., when each simulation run takes in the range of seconds.

However, the "no-bias" assumption in the linkage model (1) could be restrictive and thus often violated in real-world applications. Take the BEM calibration problem for a residential building in Texas as an example (see more details in Section 4). We use the data collected during the first 20 days of July in 2014 as a training set to calibrate the BEM parameters. We use BO for minimizing the loss function in (2). We choose BO, because it is intended to find the global minimum of the loss function when data is generated from a black-box computer model (Shahriari et al. 2015, Frazier 2018), such as the BEM. More details about BO will be discussed in Section 3.3. After calibrating $\boldsymbol{\theta}$, we check whether the bias between $y(\mathbf{x}_t)$ and $\eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}})$ exists and if it exists, how the bias pattern looks like.

Figures 1-(a) and (b) depict the actual and simulated electricity consumption patterns at each hour index $t = 1, \ldots, 480$ and every 24 hours from 1 a.m. to midnight each day, respectively. Although the simulated electricity consumption pattern $\eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}})$ mimics the actual pattern $y(\mathbf{x}_t)$ relatively well, some discrepancies are still observed between the two patterns. To examine discrepancies more specifically, we calculate residuals and plot them at each t. Let $R(\mathbf{x}_t)$ denote the residual as $R(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}})$ at time t. Figure 2-(a) clearly displays a cyclic residual pattern from $t = 1$ to 480. The residuals at a 24-hour interval, which are shown in Figure 2-(b), present the "positive-negative-positive" pattern with the "decrease-increase-slightly decrease-increase" behavior from 1 a.m. to midnight. In general, the residual $R(\mathbf{x}_t)$ tends to be positive, i.e., $y(\mathbf{x}_t) > \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}})$, during 1 a.m. to 6 a.m. and 6 p.m. to midnight, as depicted in the blue box plots in Figure 2-(c), indicating that the BEM underestimates actual electricity demands. On the contrary, it tends to overestimate actual electricity demands from 7 a.m. to 5 p.m. as shown in the red box plots in Figure 2-(c). From the observations, it becomes evident that certain daily patterns persist over time. It shows that there is a systematic bias inherent in the BEM simulator, implying that $E[y(\mathbf{x}_t)] \neq \eta(\mathbf{x}_t; \boldsymbol{\theta}^*)$ for some t.



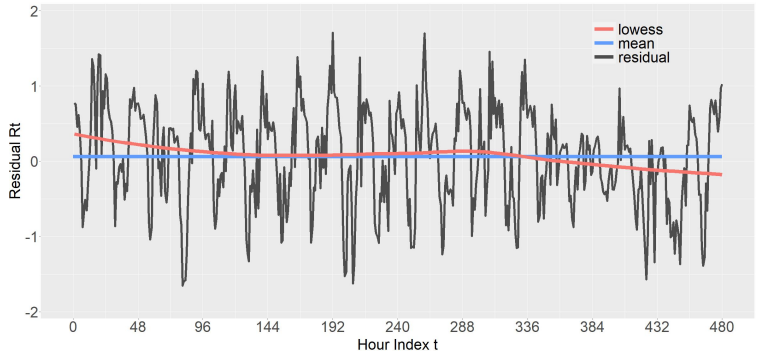
(a) Electricity consumption against hour index.

(b) Electricity consumption against every 24 hours.

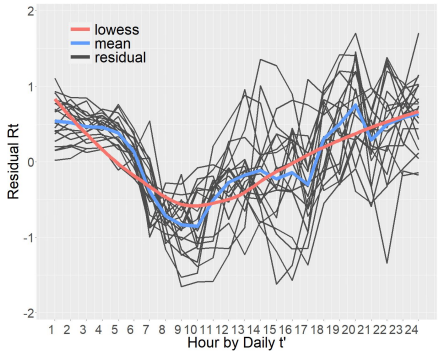
Figure 1: Comparison between the actual electricity consumption and BEM simulation outputs.

Furthermore, the autocorrelation functions (ACFs) of the residuals in Figure 3 confirm that severe temporal correlation exists within the residual time series $\{R(\mathbf{x}_t)\}_{t=1}^{T}$. Note that the ACF values that fall beyond the two horizontal dotted boundaries indicate correlated residuals. In Figure 3, we observe several ACF values beyond the boundaries, implying that the residuals are correlated with one another. Thus, the "independent" assumption in the linkage model (1), which is one of the core assumptions about the errors in $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$, is violated.

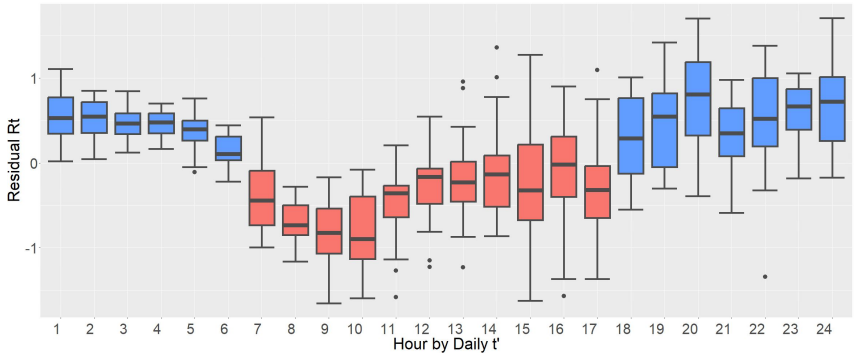
Accordingly, since ϵ_t's are not independent, the problem formulation in (2) has limitations. It implies that its optimizer $\hat{\boldsymbol{\theta}}$ is no longer the ML estimator in a statistical sense, thus one cannot use the asymptotic properties of the ML estimator (see Section 3.4) when quantifying uncertainties for the calibrated parameters $\hat{\boldsymbol{\theta}}$. To address these limitations, a bias-corrected calibration approach that adjusts the systematic discrepancy is needed. We will subsequently discuss the bias-correction procedure in more detail in the next section.



(a) Residual against hour index.



(b) Residual against every 24 hours.



(c) Box plots of residuals for each hour. The blue (resp., red) box plots indicate the residual distributions where the BEM underestimates (resp., overestimates) actual energy demands.

Figure 2: Residual plots against hour index.

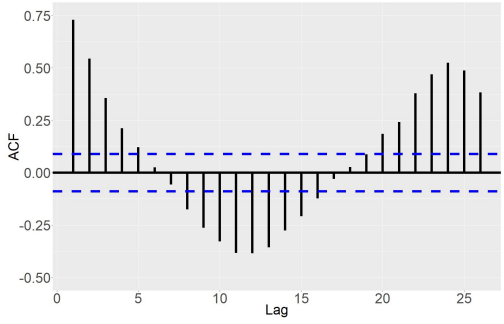


Figure 3: ACFs of the residuals.

3. Methodology: Bias-Corrected Iteratively Reweighted Least Squares Method

This section formulates the BEM calibration problem that explicitly models the bias. We also introduce heterogeneous weights into a loss function to mitigate heteroskedasticity. Specifically, Section 3.1 analyzes the bias pattern present in the BEM and discusses how to capture it using a time-series model. Section 3.2 examines heteroskedasticity in the electricity loads and formulates weighted least squares using the heterogeneous weights. Then we propose a new algorithm for the iterative refinement of weights, employing the IRLS method in Section 3.3. In Section 3.4, we show how to construct confidence intervals for the calibrated parameters. Section 3.5 discusses potential extensions to both bias correction and heterogeneous variance reduction approaches.

3.1. Debiasing the BEM outputs

Suppose $\delta(\mathbf{x}_t)$ denotes a systematic discrepancy between $y(\mathbf{x}_t)$ and $\eta(\mathbf{x}_t;\boldsymbol{\theta}^*)$ for all t. We consider the extended form of linear linkage model that incorporates $\delta(\mathbf{x}_t)$ into the previously discussed linkage model (1) as follows (Kennedy and O'Hagan 2001):

$$y(\mathbf{x}_t) = \eta(\mathbf{x}_t;\boldsymbol{\theta}^*) + \delta(\mathbf{x}_t) + \epsilon_t, \quad \forall t = 1,\dots,T, \tag{3}$$

where $\epsilon_t \overset{\text{iid}}{\sim} N(0,\sigma^2)$ holds. Later, we will see that the error assumption regarding the *identical* distribution (or constant variance σ^2) does not hold in the BEM calibration. Instead, $\epsilon_t \overset{\text{ind}}{\sim} N(0,\sigma_t^2)$ is a more valid assumption, where "ind" means "independently distributed." We will address how to characterize the varying variance σ_t^2 in Section 3.2.

When field observations and computer model outputs are scarce, Bayesian calibration (Kennedy and O'Hagan 2001) suggests modeling $\eta(\mathbf{x}_t;\boldsymbol{\theta}^*)$ and $\delta(\mathbf{x}_t)$ using surrogates, typically Gaussian processes (GPs), at the pre-designed inputs $\mathbf{x}_t$ and $\boldsymbol{\theta}$. It places a prior distribution on each parameter and explores the posteriors using Markov chain Monte Carlo (MCMC). The clear benefit of employing this Bayesian approach is to offer uncertainty quantification capabilities in the Bayesian inference framework. However, its application is usually restricted to a small amount of data with low-dimensional parameters due to the MCMC procedure's heavy computational overhead, as discussed in Section 1. It has been shown that with a large-size dataset, a frequentist approach is more useful (Liu et al. 2021, Jeong et al. 2023). However, to the best of our knowledge, the frequentist approach does not take the bias into account in the formulation (3).

In this study, our goal is to make $\eta(\mathbf{x}_t;\boldsymbol{\theta}^*) + \delta(\mathbf{x}_t)$ resemble $y(\mathbf{x}_t)$ with the judiciously modeled bias term $\delta(\mathbf{x}_t)$, so that it accurately represents the electricity demands, along with improved uncertainty quantification capabilities. We now introduce a new modeling approach to represent the biases $\{\delta(\mathbf{x}_t)\}_{t=1}^T$ using the time-series residuals $\{R(\mathbf{x}_t)\}_{t=1}^T$, in order to capture the daily cyclic (or periodic) pattern and address the temporal correlation, through the following equation.

$$R(\mathbf{x}_t) = \delta(\mathbf{x}_t) + \epsilon_t, \quad \forall t = 1,\dots,T, \tag{4}$$

where $\epsilon_t \overset{\text{iid}}{\sim} N(0,\sigma^2)$ or $\epsilon_t \overset{\text{ind}}{\sim} N(0,\sigma_t^2)$.

In general, there are a few major modeling approaches that can be considered for a time series: parametric, semi-parametric, and nonparametric approaches. The family of parametric models includes the usual time-series models such as autoregressive integrated moving average (ARIMA) (Shumway and Stoffer 2017), whereas the semiparametric and nonparametric approaches contain a relatively wide range of models such as GPs (Rasmussen and Williams 2006), splines (Hastie et al. 2009, Lee et al. 2013), neural networks including long short-term memory networks (Hochreiter and Schmidhuber 1997), etc. In this study, we utilize the parametric approach because the pattern we wish to capture is relatively consistent over time, and thus, this type of model can be easily generalized to new data that exhibit a similar pattern to the training data.

Among the family of ARIMA models, we employ the multiplicative *seasonal* autoregressive integrated moving average (SARIMA) model of a period of 24, denoted by ARIMA$(p,d,q)\times(P,D,Q)_{24}$, since the time series $\{R(\mathbf{x}_t)\}_{t=1}^T$ exhibit daily periodicity and nonstationarity, e.g., the mean value function μ_t of the residuals is not constant and depends on time t, as previously shown in Figure 2. Here, the model parameters p, d, and q are non-negative integers, with p being the order (i.e., number of time lags) of the autoregressive (AR) model, d being the degree of differencing, and q being the order of the moving average (MA) model. The uppercase letters P, D, and Q follow a similar analogy with respect to the seasonal components. Then ARIMA$(p,d,q)\times(P,D,Q)_{24}$ is expressed by

$$\Phi(B^{24})\phi(B)\nabla_{24}^D\nabla^d R_t = \alpha + \Psi(B^{24})\psi(B)\epsilon_t, \quad \forall t = 1,\dots,T, \tag{5}$$

where α is a scalar and $R_t := R(\mathbf{x}_t)$. Here, $\phi(B)$ and $\psi(B)$ are, respectively, the ordinary AR and MA operators of order p and q, defined by

$$\begin{aligned}\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,\\ \psi(B) &= 1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_q B^q,\end{aligned} \tag{6}$$

and $\Phi(B^{24})$ and $\Psi(B^{24})$ are the seasonal AR and MA operators of order P and Q, described by

$$\begin{aligned}\Phi(B^{24}) &= 1 - \Phi_1 B^{24} - \Phi_2 B^{48} - \cdots - \Phi_P B^{24P},\\ \Psi(B^{24}) &= 1 + \Psi_1 B^{24} + \Psi_2 B^{48} + \cdots + \Psi_Q B^{24Q},\end{aligned} \tag{7}$$

respectively. Additionally, $\nabla^d = (1-B)^d$ and $\nabla_{24}^D = (1-B^{24})^D$ are the ordinary and seasonal difference operators, respectively, where B denotes the backshift operator exemplified by $BR_t = R_{t-1}$.

The combination of the model parameters (p, d, q) and (P, D, Q) can be determined using the information criteria, such as AIC (Akaike 1974) and BIC (Schwarz 1978). We can find the model order that provides the lowest AIC or BIC criterion among the fitted SARIMA models. Both AIC and BIC suggest assessing the goodness of fit for the time-series model by balancing the fitting error with the model complexity represented by the number of parameters in the model. The distinction between the two criteria lies in the extent to which they penalize model complexity. Specifically, AIC is less stringent in penalizing model complexity, often leading to the selection of a larger-order model compared to BIC (Shumway and Stoffer 2017). That is, AIC may be preferred when aiming for a more flexible model and when there is a belief that a more complex model could capture important patterns in the data, while BIC generally tends to favor more parsimonious models. There is some debate surrounding the comparative advantages of these two criteria, yet both AIC and BIC have been widely employed in the literature without specific preference (Faraway 2014). In our implementation with 10 experiments (Note: Section 4 will discuss implementation settings in more detail), the same SARIMA model is selected by both AIC and BIC in five out of ten training sets. The remaining five scenarios exhibit minimal discrepancies, with a maximum of one model degree. For example, BIC selects ARIMA$(1,0,0)\times(0,1,1)_{24}$ in some experiments, when AIC favors ARIMA$(1,0,1)\times(0,1,1)_{24}$. Even though we use AIC for model selection, one can also employ BIC when seeking to select a more parsimonious model.

One could identify the degrees of difference operators that provide a relatively stationary series, followed by finding the appropriate orders for AR and MA components to fit the resulting residual series. Thus, the bias model for $\delta(\mathbf{x}_t)$ will eventually be $\delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}}) = \hat{R}(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$ by explicitly specifying the estimated model parameters $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Psi}})$, where $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)$, $\hat{\boldsymbol{\psi}} = (\hat{\psi}_1, \ldots, \hat{\psi}_q)$, $\hat{\boldsymbol{\Phi}} = (\hat{\Phi}_1, \ldots, \hat{\Phi}_P)$, and $\hat{\boldsymbol{\Psi}} = (\hat{\Psi}_1, \ldots, \hat{\Psi}_Q)$. In this example of our case study, we select the ARIMA$(1,0,0)\times(1,0,1)_{24}$ model for $\{R(\mathbf{x}_t)\}_{t=1}^T$, because (p,d,q)=$(1,0,0)$ and (P,D,Q)=$(1,0,1)$ give the lowest AIC value among the different SARIMA models. With $\alpha = 0$, $\phi_1 = \phi$, $\Phi_1 = \Phi$, and $\psi_1 = \psi$, the time-series model for $\{R(\mathbf{x}_t)\}_{t=1}^T$ becomes

$$(1-\Phi B^{24})(1-\phi B)R_t = (1+\Psi B^{24})\epsilon_t, \tag{8}$$

or equivalently, in difference equation form,

$$R_t = \phi R_{t-1} + \Phi R_{t-24} - \phi\Phi R_{t-25} + \epsilon_t + \Psi\epsilon_{t-24}. \tag{9}$$

Hence, the resulting bias model for $\delta(\mathbf{x}_t)$ becomes $\delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}}) = \hat{R}(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\beta}} = (\hat{\phi}, \hat{\Phi}, \hat{\Psi})$.

With the observed bias pattern in the BEM simulator, which is captured by $\delta(\mathbf{x}_t)$, we consider the following new loss function to calibrate the BEM parameters $\boldsymbol{\theta}$ and, at the same time, estimate $\boldsymbol{\beta}$.

$$\min_{(\boldsymbol{\theta},\boldsymbol{\beta})\in\boldsymbol{\Theta}\times\boldsymbol{\Omega}} \frac{1}{T}\sum_{t=1}^{T}(y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2. \tag{10}$$

Figure 4 shows the ACFs with this new formulation. The temporal correlation is significantly reduced by the bias-correction procedure, and model residuals (see the definition in Section 3.2) appear to be uncorrelated because the values of ACFs are within the dotted boundaries in Figure 4.

3.2. Handling Heteroskedasticity

Let us call $r(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$ as the *model residual* (recall that we denote $R(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}})$ as the *residual* to differentiate with the model residual). While the new loss function in (10) helps debias the BEM outputs and address the temporal correlation, Figure 5 (or Figure 6) still indicates that the model residual (or squared model residual) exhibits the heterogeneous variance over time t. In particular, the box plots in Figure 5-(c) and Figure 6-(c) describe the large variance of both model and squared model residuals at 2 to 5 p.m., 7 to 8 p.m., and 10 p.m., indicated by the wide interquartile ranges of the red box plots. That is, the model residual generally shows a large variance from 2 p.m. to 10 p.m., when occupants' behavior is typically stochastic (De Wilde 2014, Kim et al. 2017). It implies that the assumption of the constant error variance in $\epsilon_t \overset{\text{iid}}{\sim} N(0,\sigma^2)$ in (3) is violated. More specifically, $Var(\epsilon_t)$ is not constant, i.e., $Var(\epsilon_t) \neq \sigma^2$. In other words, the model residual time series $\{r(\mathbf{x}_t)\}_{t=1}^T$ shows heteroskedasticity over time t so they are not identically distributed.

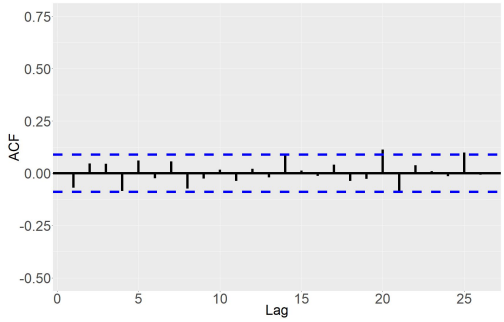
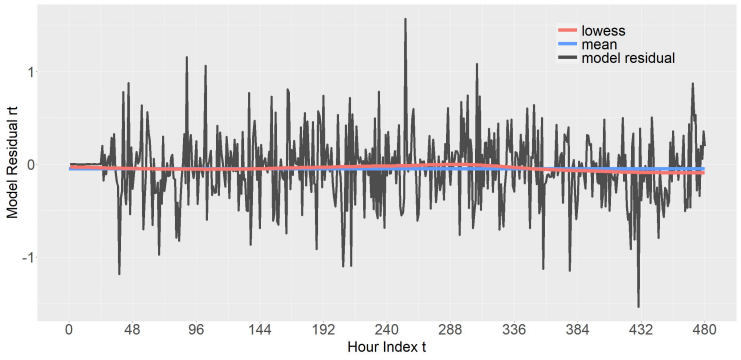
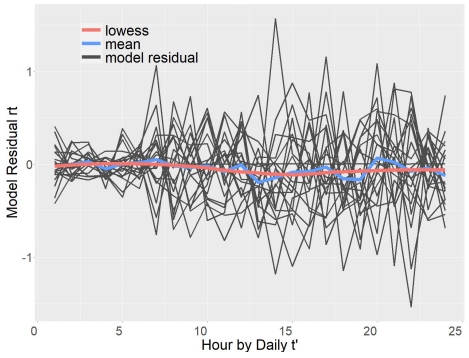


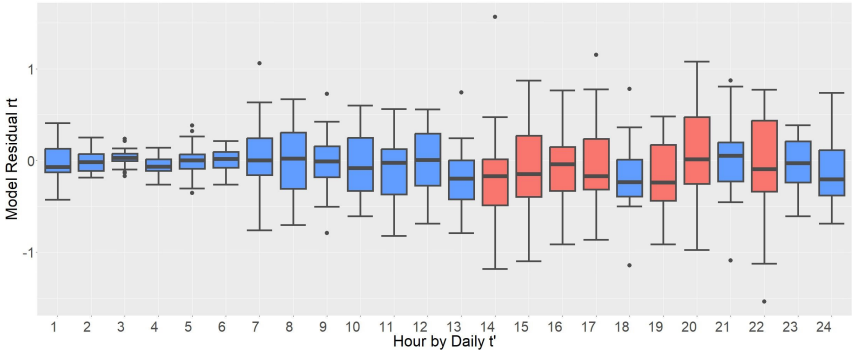
Figure 4: ACFs of the model residuals $r(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t; \hat{\boldsymbol{\beta}})$.



(a) Model residual against hour index.



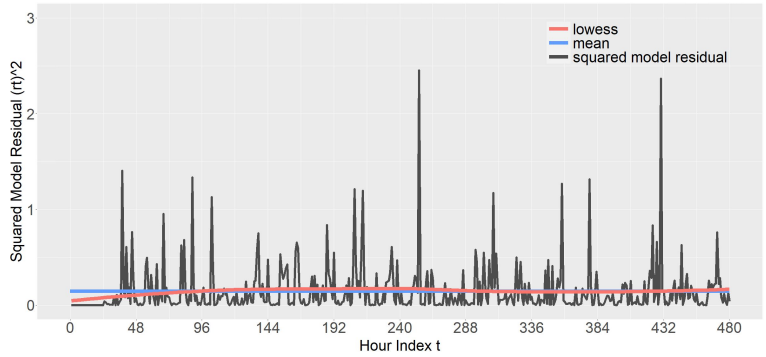
(b) Model residual against every 24 hours.



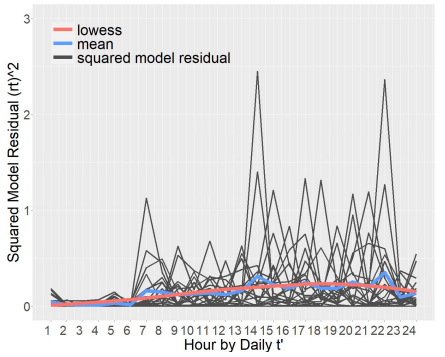
(c) Box plots of model residuals for each hour. The red box plots depict the distributions of the model residuals that show large variances.

Figure 5: Plots of model residual $r(\mathbf{x}_t)$ with $r(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t; \hat{\boldsymbol{\beta}})$ against hour index.

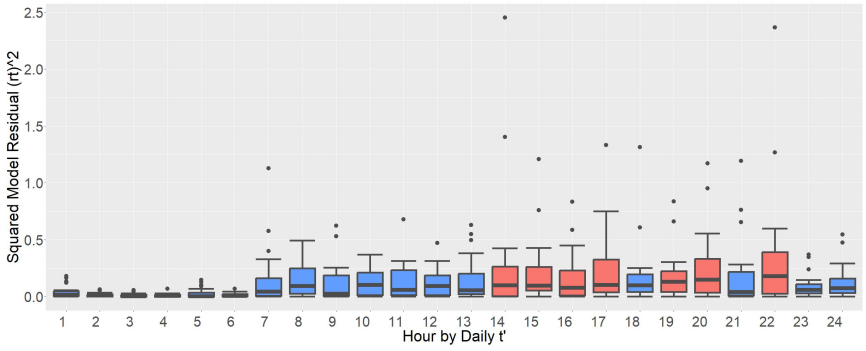
Therefore, we can conclude that ϵ_t's independently follow a normal distribution with mean 0 and variance σ_t^2, meaning that variances vary over time t, denoted by $\epsilon_t \overset{\text{ind}}{\sim} N(0, \sigma_t^2)$. It should be noted that when the variance shows heteroskedasticity, the estimates $\hat{\boldsymbol{\theta}}$ as a solution to (10) are no longer ML estimates and thus no longer enjoy ML properties, such as asymptotic normality. We will discuss how to address the heteroskedasticity issue in the subsequent discussion.



(a) Squared model residual against hour index.



(b) Squared model residual against every 24 hours.



(c) Box plots of squared model residuals for each hour. The red box plots depict the distributions of the squared model residuals that show large variances.

Figure 6: Plots of squared model residual $\{r(\mathbf{x}_t)\}^2$ with $r(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$ against hour index.

3.2.1. Mitigating Heteroskedasticity with Weights

With a heterogeneous random error $\epsilon_t \overset{\text{ind}}{\sim} N(0,\sigma_t^2)$, where σ_t^2 varies over time t, the regression model (3) can be extended as

$$y(\mathbf{x}_t) \overset{\text{ind}}{\sim} N\left(\eta(\mathbf{x}_t;\boldsymbol{\theta}) + \delta(\mathbf{x}_t;\boldsymbol{\beta}), \sigma_t^2\right). \tag{11}$$

Since $y(\mathbf{x}_t)$ independently follows the normal distribution for $t = 1,\ldots,T$, the likelihood function is as follows:

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta}|y(\mathbf{x}_{1:T})) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{(y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2}{2\sigma_t^2}\right]. \tag{12}$$

Let us take the logarithm on both side in (12). Then we have the following log-likelihood function.

$$\begin{aligned}
\ell(\boldsymbol{\theta},\boldsymbol{\beta}|y(\mathbf{x}_{1:T})) &= \log\left\{\prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{(y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2}{2\sigma_t^2}\right]\right\} \\
&= \log\left\{\prod_{t=1}^{T} (2\pi)^{-\frac{1}{2}} (\sigma_t^2)^{-\frac{1}{2}} \exp\left[-\frac{(y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2}{2\sigma_t^2}\right]\right\} \\
&= -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\sigma_t^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{(y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2}{\sigma_t^2}.
\end{aligned} \tag{13}$$

Assuming that σ_t^2 is known, the first two terms of the last equation in (13) are constant. Thus, maximizing the function (13) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is equivalent to minimizing the last term in (13), leading to minimizing the

following WLS loss function.

$$\min_{(\boldsymbol{\theta},\boldsymbol{\beta})\in\boldsymbol{\Theta}\times\boldsymbol{\Omega}} F(\boldsymbol{\theta},\boldsymbol{\beta}) := \frac{1}{T}\sum_{t=1}^{T} w_t(y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2, \tag{14}$$

where $w_t = 1/\sigma_t^2$. One can refer to Casella and Berger (2002), Kutner et al. (2005), and Faraway (2014) for more details. By solving this WLS, we can get the estimates $\hat{\boldsymbol{\theta}}$ as calibrated parameter values.

The WLS loss function in (14) has important implications when heteroskedasticity is observed, as in Figures 6-(b) and (c). The WLS gives different degrees of importance, or weights, to each data point. This is achieved by putting bigger weights when the variance of the model residuals is small, that is, when the uncertainty is small, because the points are close to the true mean. On the contrary, the WLS gives smaller weights when the variance is large since the points could largely deviate from the mean.

The challenge is that the true variance σ_t^2 is unknown in practice. Therefore, we need to estimate σ_t^2, which we will discuss further in the next section within the context of the BEM calibration.

3.2.2. Estimating Weights

With the model residual $r(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$, let $s(\mathbf{x}_t)$ denote the squared model residual, i.e., $s(\mathbf{x}_t) = \{r(\mathbf{x}_t)\}^2$ for all $t = 1,\ldots,T$. In Figures 6-(b) and (c), $\{s(\mathbf{x}_t)\}_{t=1}^{T}$ mainly fluctuates from 2 p.m. to 10 p.m. on a daily basis. Suppose we have D days of training data to calibrate the BEM parameters. To represent the heterogeneous variance pattern over time in a day, we split the time series of D days of data into D segments by 24 hours from 1 a.m. to midnight. Then we re-organize them with redefined time stamps $t' = 1,\ldots,24$ in order. Thus, we have newly indexed D day-longitudinal data of $24D$ data points (i.e., 24 hours $\times$ D days), denoted by $\mathcal{D}' = \{(t', s_j(t'))\}_{t'=1,j=1}^{t'=24,j=D}$. Note that t' denotes a specific hour of each day. In this example, we set $D = 20$.

Utilizing the fact that $E[s_j(t')] = \sigma_{t'}^2$, we can fit a regression model with $\mathcal{D}'$ to estimate the variance function σ_t^2. If the pattern of $s_j(t')$ over $t' = 1,\ldots,24$ is relatively simple, a parametric model can be employed, similar to the bias model discussed in Section 3.1. Otherwise, either a semiparametric model or a nonparametric model can be employed, especially when the pattern is not simple and thus specifying its functional form *a priori* is not desirable. Here, we employ a semiparametric model because of its flexibility. In our implementation, we specifically utilize smoothing spline (Hastie et al. 2009, Lee et al. 2013). Let $\varphi(t')$ denote the function we wish to estimate, i.e., $\varphi(t') = E[s_j(t')]$ for all t' and j. We can obtain its estimate by solving the following minimization problem.

$$\hat{\varphi} = \arg\min_{\varphi} \sum_{t'=1}^{24}\sum_{j=1}^{D}(s_j(t') - \varphi(t'))^2 + \lambda\int(\varphi''(u))^2 du, \tag{15}$$

where λ is a smoothing parameter. When λ is small, $\hat{\varphi}$ will be wiggly, whereas with large λ, $\hat{\varphi}$ will be smooth, approaching to the least squares line fit. We can choose λ using cross-validation or generalized cross-validation (Hastie et al. 2009).

The solution to (15) can be expressed by an explicit, finite-dimensional, and unique minimizer which is a natural cubic spline with knots at each t' for $t' = 1,\ldots,24$ as follows:

$$\hat{\varphi}(t') = \sum_{t'=1}^{24} N_{t'}(t')\gamma_{t'}, \tag{16}$$

where $N_{t'}(t')$ is a 24-dimensional set of basis functions for representing the family of natural splines and $\gamma_{t'}$ is the model parameters to be estimated. With this model $\hat{\varphi}(t')$, we estimate the variance function σ_t^2 by $\hat{\sigma}_t^2 = (\hat{\varphi}(t'),\ldots,\hat{\varphi}(t'))$ for $t = 1,\ldots,24D$, by stacking the D time series $\hat{\varphi}(t')$ for $t' = 1,\ldots,24$ in order. Then the weights w_t in (14) can be replaced by $w_t = 1/\hat{\sigma}_t^2$ to obtain the WLS estimate $\hat{\boldsymbol{\theta}}$.

3.3. Parameter Estimation through Iterative Refinement of Weights

As we have seen thus far, the estimate $\hat{\sigma}_t^2$ depends on the initially estimated $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ with the regression function $\eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}}) + \delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$ for $t = 1,\ldots,T$. However, considering heteroskedasticity within the WLS formulation can also

change both estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, consequently affecting the regression function. This observation suggests an iterative estimation approach, where we cyclically estimate $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\{\sigma_t^2\}_{t=1}^T$, each of which improves the other. Specifically, at kth iteration, we estimate $\boldsymbol{\theta}^{k+1}$ and $\boldsymbol{\beta}^{k+1}$. Then, given $\boldsymbol{\theta}^{k+1}$ and $\boldsymbol{\beta}^{k+1}$, we obtain an estimate for the variance function $\{(\sigma_t^{k+1})^2\}_{t=1}^T$. In the next iteration, we re-estimate $\boldsymbol{\theta}^{k+2}$ and $\boldsymbol{\beta}^{k+2}$ and get the updated regression function $\eta(\mathbf{x}_t;\boldsymbol{\theta}^{k+2}) + \delta(\mathbf{x}_t;\boldsymbol{\beta}^{k+2})$ by using WLS with the weights inversely proportional to the previously estimated variance, i.e., $w_t^{k+1} = 1/(\sigma_t^{k+1})^2$ for all t. As this re-estimation generally alters the other estimates, it also affects the residuals.

This iterative procedure continues until the change in either consecutive parameter values or loss function values becomes sufficiently small, or the available computational budget, e.g., number of simulations, is exhausted. Algorithm 1 summarizes the IRLS procedure for parameter calibration with the bias-correction component. We call Algorithm 1 the bias-corrected iteratively reweighted least squares method, abbreviated by "deBias-IRLS" hereafter.

Algorithm 1 Bias-Corrected Iteratively Reweighted Least Squares Method for Parameter Calibration (deBias-IRLS)

Input: field data $\mathcal{D}_T = \{(\mathbf{x}_t, y(\mathbf{x}_t))\}_{t=1}^T$.
Initialize the model bias $\delta(\mathbf{x}_t;\boldsymbol{\beta}^1) = 0$ and $w_t^1 = 1$, $\forall t = 1,\ldots,T$.
for $k = 1, 2, \ldots, K_{\max}$ **do**
  Step 1 (update $\boldsymbol{\theta}$):
    Given $\delta(\mathbf{x}_t;\boldsymbol{\beta}^k)$, perform BO using Algorithm 2 to obtain the $(k+1)$th iterate of parameters $\boldsymbol{\theta}^{k+1}$, i.e.,

$$\boldsymbol{\theta}^{k+1} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\, F(\boldsymbol{\theta},\boldsymbol{\beta}^k,\boldsymbol{w}^k) := \frac{1}{T}\sum_{t=1}^{T} w_t^k (y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}^k))^2. \tag{17}$$

    Calculate the residual as $R_t^{k+1} := R^{k+1}(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}^{k+1})$, $\forall t$.
  Step 2 (update $\boldsymbol{\beta}$):
    Given $\eta(\mathbf{x}_t;\boldsymbol{\theta}^{k+1})$, fit a time-series model $\delta(\cdot)$ to $\{R_t^{k+1}\}_{t=1}^T$ to obtain $\boldsymbol{\beta}^{k+1}$, i.e.,

$$\boldsymbol{\beta}^{k+1} \leftarrow \underset{\boldsymbol{\beta}\in\boldsymbol{\Omega}}{\arg\min}\, F(\boldsymbol{\theta}^{k+1},\boldsymbol{\beta},\boldsymbol{w}^k) := \frac{1}{T}\sum_{t=1}^{T} w_t^k (y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}^{k+1}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2. \tag{18}$$

    Calculate the model residual as $r_t^{k+1} := r^{k+1}(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}^{k+1}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}^{k+1})$, $\forall t$.
  Step 3 (update $\boldsymbol{w}$):
    Fit a regression model $\{(\sigma_t^{k+1})^2\}_{t=1}^T$ on $\{(r_t^{k+1})^2\}_{t=1}^T$ to get

$$w_t^{k+1} = \frac{1}{(\sigma_t^{k+1})^2}, \quad \forall t. \tag{19}$$

  If a termination condition holds, **then** break the loop.
end for
Output: calibrated parameters $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^k$, bias model $\delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}}) = \delta(\mathbf{x}_t;\boldsymbol{\beta}^k)$, and weights $\hat{w}_t = w_t^k$ with $\hat{\sigma}_t^2 = (\sigma_t^k)^2$, $\forall t$.

In Step 1 of Algorithm 1, we employ BO to get $\boldsymbol{\theta}^{k+1}$. Note that BO is completely different method from Bayesian calibration. For Bayesian calibration, one can refer to Sections 2, 3.1, and 4.1. We choose BO due to its strong capability to handle a black-box and derivative-free optimization problem, even though other optimization techniques such as gradient descent and second-order optimization methods can also be used with gradient approximation. We briefly explain the concept of BO here. One can find more thorough discussions in Frazier (2018) and Shahriari et al. (2015). BO is a global optimization method to minimize the loss function $F(\cdot)$, which is (i) expensive-to-evaluate, (ii) black-box, and (iii) derivative-free. BO works well in our setting, because evaluating the value of $\eta(\cdot)$ in $F(\cdot)$ by running the BEM is not instantaneous. It takes approximately 1–2 minutes for one year-long BEM simulation. Moreover, $\eta(\cdot)$ lacks an explicit expression due to intricate mathematical functions within the simulator, precluding the availability of first- and second-derivative information in general. Thus, we believe that BO, as an optimization tool, is an adequate choice for the BEM parameter calibration.

Specifically, BO starts with constructing a GP for $F(\cdot)$ with an initial space-filling design of N_0 points (or parameter

settings in our problem context) such as Latin hypercube design (LHD) (McKay et al. 1979), maximin distance design (Jones et al. 1998), and maximin LHD (Morris and Mitchell 1995). Given the GP, an acquisition function $\mathcal{A}(\cdot)$, such as the expected improvement (EI) (Močkus 1975, Jones et al. 1998), upper/lower confidence bounds (Srinivas et al. 2010), and knowledge gradient (Frazier et al. 2009), is maximized to find the next design point by striking a balance between exploration and exploitation. Then $F(\cdot)$ is evaluated at this design point, and the GP is updated accordingly. This procedure continues until the algorithm converges or the maximum number of simulation budgets $N_{\max}$ is exhausted. Finally, we get the minimizer $\boldsymbol{\theta}_{\min}$ that shows the lowest loss function value thus far. In our implementation, we utilize the maximin LHD for the space-filling design and the EI for the acquisition function. The EI is defined as $\mathrm{EI}(\boldsymbol{\theta}|\mathcal{D}) = E[\max(F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}_{\min}), 0)]$ with all available data $\mathcal{D}$, which can be evaluated using the closed-form solution (Jones et al. 1998). We implement our procedure with the statistical software R (R Core Team 2021). Among several GP and BO packages available in R, we use `DiceKriging` and `DiceOptim` (Roustant et al. 2012) due to their wide popularity. We summarize the BO procedure in Algorithm 2.

Additionally, it is worth noting that since the regression function consists of two terms, $\eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}})$ and $\delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$, performing many iterations poses the risk of potentially losing some important aspects of the pattern that should be captured by the computer model $\eta(\cdot)$. In fact, there is a possibility that this pattern could be assimilated into the bias term $\delta(\cdot)$, which is undesirable, because $\delta(\cdot)$ should serve as a supplementary term to correct the possible bias. This aligns with the identifiability issue between $\eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}})$ and $\delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$, and we admit that our proposed method does not address this identifiability issue. Devising a procedure that uniquely estimates $\eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}})$ and $\delta(\mathbf{x}_t;\hat{\boldsymbol{\beta}})$ is beyond the scope of this study. However, considering the importance of $\eta(\cdot)$ over $\delta(\cdot)$, the practical remedy is to terminate the procedure within a small number of iterations, e.g., 5 iterations.

Algorithm 2 Bayesian Optimization (BO)

1: **Input:** N_0, $N_{\max}$ ($> N_0$), and an acquisition function $\mathcal{A}(\cdot)$.
2: Evaluate $F(\cdot)$ at N_0 points of $\boldsymbol{\theta}$, generated by a space-filling experimental design. Obtain the initial points $\mathcal{D}_{N_0} = \{\boldsymbol{\theta}_n, F(\boldsymbol{\theta}_n)\}_{n=1}^{N_0}$.
3: Place an initial GP prior on $F(\cdot)$ with $\mathcal{D}_{N_0}$ by estimating the GP hyperparameters.
4: **for** $n = N_0 + 1, \ldots, N_{\max}$ **do**
5: Obtain $\boldsymbol{\theta}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{A}(\boldsymbol{\theta}|\mathcal{D}_{n-1})$.
6: Evaluate $F(\cdot)$ at $\boldsymbol{\theta}_n$ and set $\mathcal{D}_n = \mathcal{D}_{n-1} \cup (\boldsymbol{\theta}_n, F(\boldsymbol{\theta}_n))$.
7: Update the GP posterior with $(\boldsymbol{\theta}_n, F(\boldsymbol{\theta}_n))$.
8: **end for**
9: **Output:** the point with the lowest $F(\boldsymbol{\theta})$, i.e., $\boldsymbol{\theta}_{\min}$.

3.4. Uncertainty Quantification

In this section, we discuss how to quantify the uncertainties of the calibrated parameters $\hat{\boldsymbol{\theta}}$ by constructing the confidence intervals (CIs) for $\boldsymbol{\theta}$ (Choe et al. 2018, Pan et al. 2021, Jeong et al. 2023). Let us standardize $y(\mathbf{x}_t)$ in (11) to follow the standard normal distribution $N(0, 1^2)$. Then we have

$$Z_t = \frac{y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta})}{\sigma_t} \overset{\text{iid}}{\sim} N(0, 1^2). \tag{20}$$

The standardization in (20) is useful because the ML estimates $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ with iid observations possess compelling theoretical properties, including consistency, asymptotic normality, and efficiency. Consistency tells that $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ converges in probability to the true parameters $\boldsymbol{\theta}_{\mathrm{true}}$ as $T \to \infty$, denoted by $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{true}}$. Further, asymptotic normality indicates that the estimator $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_{\mathrm{true}})$ converges in distribution to a (multivariate) normal distribution $N(\mathbf{0}, I(\boldsymbol{\theta}_{\mathrm{true}})^{-1})$ as $T \to \infty$, or concisely, $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_{\mathrm{true}}) \xrightarrow{d} N(\mathbf{0}, I(\boldsymbol{\theta}_{\mathrm{true}})^{-1})$, where $I(\boldsymbol{\theta}_{\mathrm{true}})$ is an expected Fisher information matrix. Also, it is known that $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is asymptotically efficient, that is, $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ attains its Cramér-Rao lower bound for large samples T (Casella and Berger 2002).

For uncertainty quantification, we use the ML estimator's asymptotic properties to construct CIs for the parameters (Jeong et al. 2023). Consider the expected Fisher information matrix represented by

$$I(\boldsymbol{\theta}_{\text{true}}) = E\left(\frac{\partial \ell_1(\boldsymbol{\theta}|y(\mathbf{x}))}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell_1(\boldsymbol{\theta}|y(\mathbf{x}))}{\partial \boldsymbol{\theta}}\right)^{\mathsf{T}}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{true}}}, \tag{21}$$

where $\ell_1(\boldsymbol{\theta}|y(\mathbf{x}))$ is a log-likelihood function at a single observation $y(\mathbf{x})$ at $\mathbf{x}$. With $\mathbf{x} = \mathbf{x}_t$ the log-likelihood function $\ell_1(\boldsymbol{\theta}|y(\mathbf{x}_t))$ is

$$\ell_1(\boldsymbol{\theta}|y(\mathbf{x}_t)) = -\frac{1}{2}\log 2\pi - \frac{1}{2\sigma_t^2}(y(\mathbf{x}) - \eta(\mathbf{x}_t;\boldsymbol{\theta}) - \delta(\mathbf{x}_t;\boldsymbol{\beta}))^2, \quad \forall t = 1,\ldots,T, \tag{22}$$

assuming that we know $\{\sigma_t^2\}_{t=1}^T$ and $\boldsymbol{\beta}$. In practice, we can replace these two quantities with their estimates $\{\hat{\sigma}_t^2\}_{t=1}^T$ and $\hat{\boldsymbol{\beta}}$, respectively.

The expected Fisher information matrix in (21) can be approximated by its empirical counterpart,

$$I(\hat{\boldsymbol{\theta}}_{\text{ML}}) \approx \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \ell_1(\boldsymbol{\theta}|y(\mathbf{x}_t))}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell_1(\boldsymbol{\theta}|y(\mathbf{x}_t))}{\partial \boldsymbol{\theta}}\right)^{\mathsf{T}}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{ML}}}. \tag{23}$$

Here, the black-box nature of $\eta(\cdot)$ does not allow us to analytically obtain the first-order partial derivatives of $\ell_1(\boldsymbol{\theta}|y(\mathbf{x}_t))$. Instead, we use the central finite difference (Abramowitz and Stegun 1972) to numerically attain them as follows:

$$\frac{\partial \ell_1(\boldsymbol{\theta}|y(\mathbf{x}_t))}{\partial \theta_i} \approx \frac{\ell_1(\boldsymbol{\theta} + h\boldsymbol{e}_i|y(\mathbf{x}_t)) - \ell_1(\boldsymbol{\theta} - h\boldsymbol{e}_i|y(\mathbf{x}_t))}{2h}, \tag{24}$$

for $i = 1,\ldots,P_\theta$, where $h > 0$ is small number such as 10^{-8} and $\boldsymbol{e}_i$ is a $P_\theta \times 1$ vector with its ith element being one and others zero. Then we can obtain the asymptotic $100(1-\alpha)\%$ Wald CI for each component of $\hat{\boldsymbol{\theta}}_{\text{ML}}$ as follows:

$$\hat{\theta}_{\text{ML},i} \;\pm\; z_{1-\alpha/2}\frac{1}{\sqrt{T}}\sqrt{I_{ii}^{-1}(\hat{\boldsymbol{\theta}}_{\text{ML}})}, \tag{25}$$

for $i = 1,\ldots,P_\theta$, where $\hat{\theta}_{\text{ML},i}$ is the ith component of $\hat{\boldsymbol{\theta}}_{\text{ML}}$, $z_{1-\alpha/2}$ is a critical value of the standard normal distribution, and $I_{ii}^{-1}(\cdot)$ denotes the ith diagonal entry of the inverse of the Fisher information matrix $I(\cdot)$.

3.5. Extension of deBias-IRLS

The proposed approach is flexible in capturing the bias when the bias displays a unique temporal pattern throughout a day. It is worth noting that the bias pattern does not need to be cyclical in our approach. Let us consider a case where the building energy model effectively identifies and tracks the daily periodic pattern in energy use and thus, the residual $R(\mathbf{x}_t) = y(\mathbf{x}_t) - \eta(\mathbf{x}_t;\hat{\boldsymbol{\theta}})$ does not present a consistent periodic pattern. Nevertheless, the SARIMA model in our proposed method remains valid even under this condition. This is because the model can be simplified to an ARIMA model through the elimination of the periodic components with $(P, D, Q) = (0, 0, 0)$ (Note: refer to Section 3.1 for notations in SARIMA).

Furthermore, the proposed approach can be employed for either heterogeneous or homogeneous variance patterns. Take, for instance, a scenario where the variance of electricity loads and the associated bias pattern are relatively homogeneous. Under this condition, the proposed methodology retains its applicability because WLS naturally reduces to ordinary least squares with identical weight values involved.

Despite its flexibility, our proposed methodology needs careful preliminary analysis. It operates on the presumption that the bias displays a distinct temporal pattern over a day, thereby making a time-series model an appropriate choice for a bias model. Our analysis presented in Section 2 suggests that both building energy consumption and its corresponding bias indeed exhibit a daily periodicity over time. However, in cases where bias and variance patterns show non-periodic patterns or they are related with other factors such as temperature and humidity, the bias model should be adjusted to account for such factors. In such cases, the bias model may need to comprise other types of parametric or nonparametric models beyond the time-series model. Therefore, to identify the most suitable bias model, a comprehensive exploratory data analysis must be conducted prior to the application of the proposed method. Similarly, when the variance patterns interlink with other factors, WLS should leverage appropriate regression functions for these associated factors.

4. Case Study: Electrical Energy Demand Prediction with Parameter Calibration

We assess the effectiveness of the proposed calibration approach using hourly electricity consumption data during the summer months from June to September 2014, obtained from a residential building located in the Mueller neighborhood area of Austin, Texas. To simulate the building's electricity demand, we first initialize a BEM using BEopt 2.8.0.0 (Christensen et al. 2011), which is an EnergyPlus-based software for evaluating residential building designs, by taking the specific dimensions and other relevant characteristics of the building. Here, the building is configured in a rectangular shape with dimensions of $14 \times 9\ m^2$ on the first floor and $10 \times 9\ m^2$ on the second floor.

To run the EnergyPlus BEM simulator, two input files should be established beforehand: EPW and IDF files. The EPW file stands for EnergyPlus weather file and typically includes weather information such as dry-bulb temperature, relative humidity, wind speed and direction, and atmospheric pressure near the building under study. In our analysis, they are collected from the meteorological station situated nearby the building in Mueller and compiled in the EPW file with a 1-hour resolution during the corresponding period of energy consumption data. These ambient conditions serve as observable inputs $\mathbf{x}_t$ in (3) within the context of calibration.

The IDF file is an abbreviation of the input data file and contains various information that defines the simulation setting and environment. It typically includes (i) schedules for various building operations that dictate when and how they operate during the simulation; (ii) building description such as the building's geometry as well as constructions, zones, and thermal properties; (iii) HVAC system description; (iv) information about internal loads such as occupancy schedules, lighting, and equipment loads; (v) zone conditions that describe desired thermal comfort conditions, thermostat setpoints, and HVAC control sequences for each thermal zone; (vi) material and construction properties, etc. (U.S. Department of Energy 2019). These pieces of information are characterized by simulation parameters. Users can adjust the parameters and customize the IDF file according to their modeling requirements and goals. The parameters used in this study are described in Section 4.1. With the two input files, we simulate hourly electricity consumption for the studied building using EnergyPlus 9.3.0 (U.S. Department of Energy 2019).

4.1. Implementation Settings

In the BEM calibration literature, various studies have targeted specific sets of parameters. For instance, Manfren et al. (2013) concentrated on parameters related to lighting, control and operation systems, water and air loops, air handling units, and domestic hot water. Chong et al. (2017) centered their investigation on parameters associated with walls and materials. Chong and Menberg (2018) incorporated multiple parameters related to HVAC systems, envelope (such as wall and roof) thermal characteristics, and internal load-related parameters. Kim and Park (2016) considered multiple parameters for thermal zones, fans and pumps, and plants, to name a few. In this study, we select four parameters that are considered important in the literature, as summarized in Table 1.

Table 1: List of BEM parameters and their ranges.

Symbol	Description	Unit	Default	Min	Max
θ_1	Solar transmittance	–	0.4	0	1
θ_2	Gross rated cooling COP	W/W	2.95	2	5
θ_3	Gross rated cooling capacity	W	30517	12000	60000
θ_4	Cooling supply air flow rate	m^3/s	0.77	0	1

To evaluate the predictive performance of the proposed method for simulating the electricity demands of the building, we consider multiple training and test sets (Lee and Tong 2012, Lü et al. 2015, Granderson et al. 2016). Specifically, we use actual and simulated hourly electricity consumption data over a period of 21 consecutive days as a training set, whereas as a testing set, we use the data collected over the following 10 days. We consider 10 different "training (21 consecutive days)–test (next 7 consecutive days)" pairs of data by shifting the time horizon from June to September in 2014. For example, the first dataset, `Data 1`, consists of 504 hourly electricity consumption data (21 days during Jun 1–Jun 21) for training and 168 hourly data (7 days during Jun 22–Jun 28) for testing. `Data 2` also comprises 504 hourly data (Jun 8–Jun 28) for training and 168 hourly data (Jun 29–Jul 5) for testing. Note that

the time period is shifted by 7 days to account for weekly variations. Similarly, we set eight additional pairs of data, `Data 3` to `Data 10`. We conduct 10 experiments with these 10 sets of training and test datasets to assess the proposed method.

For the BO implementation in Step 1 of Algorithm 1 (or Algorithm 2), we set the initial number of design points N_0 as 40 (=$10 \times P_\theta$) to ensure the good quality of the initial GP surrogate (Loeppky et al. 2009). The maximum simulation budget $N_{\max}$ is set to 300 in Algorithm 2. For the selection of model parameters in the SARIMA model in Step 2 of Algorithm 1, they will be fixed once the initial selection is made in order to ensure stable performance of the algorithm. Further, Algorithm 1 terminates when either (i) the Euclidean distance between the consecutive parameter values is less than 10^{-2}, (ii) the difference between the consecutive loss function values $F(\cdot)$ is less than 10^{-2}, or (iii) the maximum number of iterations is reached, e.g., we set $K_{\max} = 5$ in this case study.

For the purpose of comparing the effectiveness of the proposed approach with alternative methods, we employ three standard metrics: MSE, CVRMSE, and NMBE. The MSE serves as our primary loss function which we aim to minimize. We calculate it using the test set for out-of-sample prediction, and it quantifies the discrepancy between the measured and simulated energy consumption in unseen data.

$$MSE\ [kWh^2] = \frac{1}{T_{\text{test}}} \sum_{t=1}^{T_{\text{test}}} (y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t; \hat{\boldsymbol{\beta}}))^2, \tag{26}$$

where T_{test} denotes the number of data points in the test set, i.e., $T_{\text{test}} = 168$ in this case study.

Next, the CVRMSE (Coefficient of Variation of the Root Mean Squared Error) similarly quantifies the discrepancy between the two, but it divides the square root of MSE by the averaged measured value $\overline{y(\mathbf{x})}$ as follows:

$$CVRMSE\ [\%] = \frac{1}{\overline{y(\mathbf{x})}} \sqrt{\frac{\sum_{t=1}^{T_{\text{test}}} (y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t; \hat{\boldsymbol{\beta}}))^2}{T_{\text{test}}}} \times 100. \tag{27}$$

Both the ASHRAE Guideline 14 (ASHRAE 2002, 2014) and the protocol offered by Federal Energy Management Program (FEMP) (Webster and Bradford 2002, Webster et al. 2014), which describe the measurement and verification of BEMs, recommend that CVRMSE should not exceed 30% for hourly data to ensure the well-validated BEM.

Finally, the NMBE (Normalized Mean Bias Error) is a normalized form of the mean bias error calculated by the average discrepancies between measured and simulated data as below:

$$NMBE\ [\%] = \frac{1}{\overline{y(\mathbf{x})}} \frac{\sum_{t=1}^{T_{\text{test}}} (y(\mathbf{x}_t) - \eta(\mathbf{x}_t; \hat{\boldsymbol{\theta}}) - \delta(\mathbf{x}_t; \hat{\boldsymbol{\beta}}))}{T_{\text{test}}} \times 100. \tag{28}$$

A positive (negative) NMBE value indicates whether the BEM underestimates (overestimates) the measured energy consumption values. Typically, this metric is not employed in isolation but rather used as a supplementary criterion in conjunction with MSE and CVRMSE due to its potential for cancellation effects, i.e., large positive model residuals can be cancelled by large negative ones. ASHRAE Guideline 14 and the FEMP protocol suggest that NMBE from a well-calibrated BEM should fall within the range of ±5%.

4.2. Comparison with other alternatives

We compare the performance of deBias-IRLS with other alternatives, including ordinary least squares, Bayesian calibration, L_2 calibration, and two approaches using artificial neural networks.

(a) **Ordinary Least Squares (OLS):** This method is a straightforward approach that calibrates $\boldsymbol{\theta}$ by directly solving the minimization problem (2) without considering the bias as well as heteroskedasticity. In other words, when evaluating the OLS method, the bias term $\delta(\cdot)$ in the MSE, CVRMSE, and NMBE formula is set to 0 since OLS does not consider the bias. We employ BO to get parameter estimates $\hat{\boldsymbol{\theta}}$ with OLS.

(b) **Bayesian Calibration:** It has been a predominant approach in the calibration literature. It is usually built upon the linear linkage model (3). Unlike OLS, it considers the bias term $\delta(\cdot)$, which is usually emulated by a GP. The BEM output $\eta(\mathbf{x}_t; \boldsymbol{\theta})$ is also modeled by using a GP with pre-designed points $\mathbf{x}_t$ and $\boldsymbol{\theta}$. The parameters $\boldsymbol{\theta}$ along with hyperparameters in a kernel function of the GPs are placed by the prior distributions using domain knowledge or non-informative priors and estimated by exploring their posteriors using the MCMC procedure.

(c) **L_2 Calibration:** This calibration method takes two steps to obtain an estimator $\hat{\boldsymbol{\theta}}^{L_2}$ for the true parameter $\boldsymbol{\theta}^*$. Given the inputs and corresponding actual electricity consumption data $\{(\mathbf{x}_t, y(\mathbf{x}_t))\}_{t=1}^{T}$, it proceeds by estimating the true process $\hat{\zeta}$, where $y(\mathbf{x}) = \zeta(\mathbf{x}) + \epsilon$ with an observation error ϵ, using kernel ridge regression (Hastie et al. 2009, Byon et al. 2016) in the reproducing kernel Hilbert space. Then it constructs an emulator $\hat{\eta}(\cdot)$ for $\eta(\cdot)$ by a GP (Santner et al. 2018) using the pre-specified design points of parameters and their BEM outputs $\{((\mathbf{x}_{t'}; \boldsymbol{\theta}_{t'}), \eta(\mathbf{x}_{t'}; \boldsymbol{\theta}_{t'}))\}_{t'=1}^{T'}$. Consequently, $\boldsymbol{\theta}$ is calibrated by solving the following optimization problem:

$$\hat{\boldsymbol{\theta}}^{L_2} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \ \|\hat{\zeta}(\cdot) - \hat{\eta}(\cdot\,; \boldsymbol{\theta})\|_{L_2}. \tag{29}$$

(d) **Neural Networks I:** This approach mimics the inverse model-based calibration framework (Bhatnagar et al. 2022) using artificial neural networks. Unlike the so-called forward model-based calibration approach that maps $\boldsymbol{\theta}$ to $\eta(\cdot\,; \boldsymbol{\theta})$ to find the surrogate $\hat{\eta}(\cdot)$ in the conventional setting, the inverse model-based approach maps $\eta(\cdot\,; \boldsymbol{\theta})$ to $\boldsymbol{\theta}$ for estimating the relationship g such that $\boldsymbol{\theta} = g(\eta(\mathbf{x}; \boldsymbol{\theta}) + \delta(\mathbf{x})) + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon}$ being a P_θ-dimensional vector of random errors. To train the function g, artificial neural networks are used due to the high degree of nonlinearity of g.

(e) **Neural Networks II:** This approach is a variant of L_2 calibration, but it estimates the surrogates of $\zeta(\cdot)$ and $\eta(\cdot)$ using artificial neural networks rather than GPs.

Similar to the proposed method, we use hourly data for implementing all of these alternatives with the exception of the Bayesian technique. For Bayesian calibration, we employ daily aggregated data due to the extensive computational time required when using hourly data. Additionally, we utilize two distinct prior specifications for each parameter: uniform and Gaussian. The prior means for the uniform distributions of the normalized parameters, ranging from 0 to 1, are set to 0.5. For the Gaussian prior distributions, the prior means are set close to the parameter values obtained by deBias-IRLS, mimicking the case where we have prior knowledge, along with the standard deviation of 0.2. To explore the posterior distributions, we use the No-U-Turn Sampler (NUTS), a sampling technique based on Hamiltonian Monte Carlo, which enhances the convergence of MCMC by efficiently exploring the posterior distributions. Concerning the MCMC procedure, it is set to run for 4000 iterations with 4 chains. The first half of the samples is designated as the burn-in period, while the second half is utilized to explore the posteriors. More details about Bayesian calibration can be found in Kennedy and O'Hagan (2001) and Higdon et al. (2004) in a general setting, as well as in Chong and Menberg (2018) in the context of the BEM calibration. In this study, we do not conduct Bayesian calibration using the *hourly* data, because it takes over 2 days for each training set, and we decide to quit running the implementation due to its excessive computation time. Instead, when we use the *daily* aggregated data, it takes roughly 8–9 hours to complete each experiment. It should be noted that the computation time is also dependent on the number of MCMC iterations.

4.3. Implementation Results

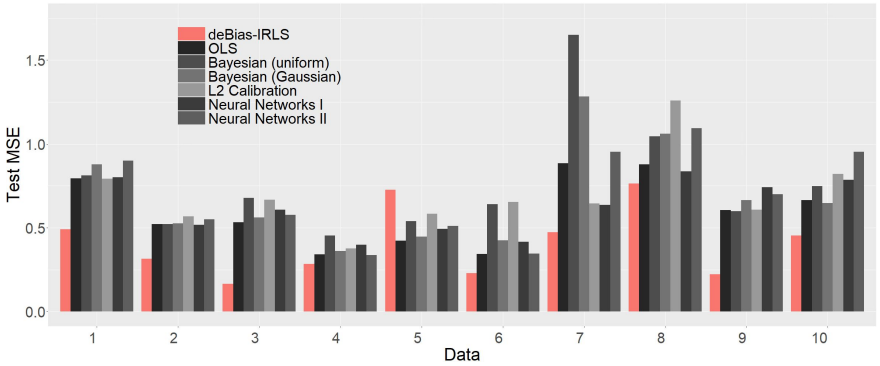
In this section, we evaluate the effectiveness of the proposed deBias-IRLS method and compare it with other alternatives in terms of calibration accuracy (or prediction accuracy) and the suitable uncertainty quantification capability. Table 2 first summarizes the comparison results based on the average values of MSEs, CVRMSEs, and NMBEs and their standard deviations, using the 10 different test sets. Clearly, deBias-IRLS achieves the lowest MSE and CVRMSE values compared to other methods on average, indicating the highest prediction accuracy for the building's electricity simulation. For more details, Figure 7 presents the comparison of MSEs and CVRMSEs of each method for 10 test sets.

Furthermore, the CVRMSE value obtained from deBias-IRLS is lower than the threshold of 30%, and the NMBE value falls within the allowable range of ±5%, both of which satisfy the guidelines from ASHRAE and FEMP. The results indicate that the BEM is well-calibrated through the proposed bias-correction procedure. Although the absolute values of NMBE from some alternative methods are slightly smaller than that of deBias-IRLS, it would be due to some cancellation effects in positive and negative residuals when calculating NMBEs. Also, NMBE is generally used as a supplementary measure alongside MSE and CVRMSE in the BEM calibration. Therefore, it is advisable to place greater emphasis on MSE and CVRMSE than NMBE when interpreting the calibration results.

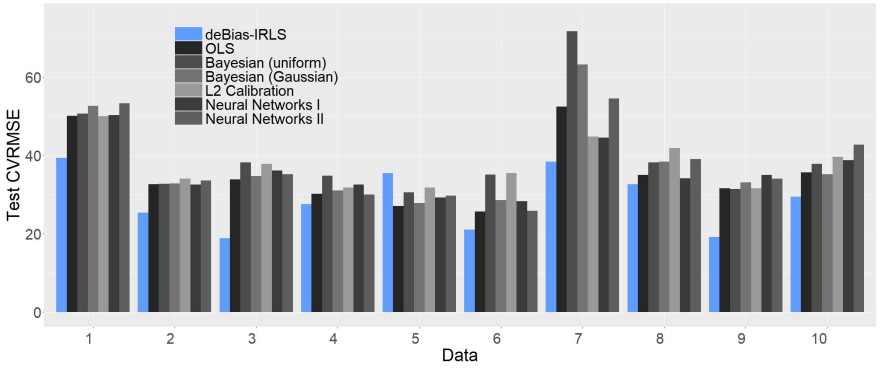
Unlike the deBias-IRLS method, OLS does not consider the systematic bias, which usually occurs in the BEM application as discussed in Section 2, so it loses its prediction capability. Moreover, the Bayesian approach exhibits

Table 2: Comparison of calibration accuracy: average MSE, CVRMSE, and NMBE in 10 test sets (Note: the values inside parentheses are standard deviations).

Method (Prior)	MSE [kWh^2]	CVRMSE [%]	NMBE [%]
deBias-IRLS	**0.412** (0.208)	**28.720** (7.662)	-2.464 (14.810)
OLS	0.598 (0.204)	35.410 (8.947)	2.752 (13.998)
Bayesian (uniform)	0.769 (0.353)	40.108 (12.441)	-9.033 (19.608)
Bayesian (Guassian)	0.685 (0.299)	37.747 (11.336)	-2.142 (19.163)
L_2 Calibration	0.697 (0.232)	37.881 (6.214)	-1.696 (13.770)
Neural Networks I	0.623 (0.163)	36.148 (6.816)	5.455 (10.054)
Neural Networks II	0.692 (0.270)	37.781 (9.734)	**0.877** (16.467)



(a) Test MSE.



(b) Test CVRMSE.

Figure 7: Comparison of calibration accuracy in terms of MSE and CVRMSE for each test set.

worse prediction performance compared to deBias-IRLS and OLS in terms of MSE and CVRMSE. Interestingly, even when the Gaussian prior is employed with prior means close to the values from the deBias-IRLS, the Bayesian approach generates higher error measures. One possible reason is that it uses daily data for training due to its excessive computing time with hourly data. Another reason of the low prediction accuracy is that the Bayesian approach utilizes only the pre-designed points in calibration to build their surrogates, whereas deBias-IRLS adaptively finds the parameter points by sequentially simulating the energy consumption output within the BO implementation. Additionally, the respective calibration objective is different in that deBias-IRLS directly minimizes the difference between the actual and simulated electricity consumption along with the bias, whereas the Bayesian calibration maximizes the likelihood with the surrogates constructed to represent the BEM and/or physical process. Also, L_2 calibration and two benchmarks using neural networks show lower calibration accuracy than deBias-IRLS. This is possibly because

L_2 calibration and two neural networks do not deal with the bias appropriately. Another reason could be that L_2 calibration and Neural Network II still construct the surrogates using the pre-designed points.

Another advantage of the deBias-IRLS method is that it provides improved uncertainty quantification capabilities once the ML estimates $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ are obtained. Recall that putting larger (or smaller) weights on less (or more) varying periods through the WLS formulation alleviates heteroskedasticity. This formulation enables us to use the asymptotic properties of the ML estimator and thus to construct the asymptotic CIs for the BEM parameters. In addition to the aforementioned alternatives, we consider another benchmark that just addresses the systematic bias through bias modeling, ignoring heteroskedasticity, in order to demonstrate the impact of mitigating heteroskedasticity. We call this additional benchmark deBias-OLS.

The performance of uncertainty quantification can be evaluated based on the CIs of the estimated parameters (Pan et al. 2021, Jeong et al. 2023). Table 3 summarizes the comparison results of the half-bandwidth of 95% CIs for each parameter using each method. Note that CIs stand for confidence intervals for OLS, deBias-IRLS, L_2 calibration, and two methods using neural networks, whereas credible intervals for Bayesian calibration. All the methods except for Bayesian calibration use the asymptotic properties of the ML to construct the confidence intervals. On the contrary, the 95% credible intervals in Bayesian calibration are derived by the intervals with the 0.025 and 0.075 quantiles of each posterior distribution. A larger bandwidth indicates a greater uncertainty in estimation. Overall, both deBias-IRLS and deBias-OLS yield narrower CIs compared to other methods. It implies that the bias correction may assist lower estimation uncertainties even if its goal is to eliminate any potential bias patterns. When compared to deBias-OLS, deBias-IRLS provides either narrower or comparable half-bandwidths of CIs, indicating more controlled uncertainty. Also, deBias-IRLS constructs narrower CIs than OLS, L_2 calibration, and Neural Networks I and II. Further, its standard deviations (see the numbers inside the parentheses) are smaller in general, indicating better robustness. Finally, it is not straightforward to compare between confidence and credible intervals, but the results suggest that deBias-IRLS leads to much narrower half-bandwidths than the Bayesian approach.

Table 3: Uncertainty quantification results: average half-bandwidth of the 95% CI for each BEM parameter in 10 test sets (Note: the values inside parentheses are standard deviations).

Method (Prior)	θ_1	θ_2	θ_3	θ_4
deBias-IRLS	0.045 (0.071)	0.032 (0.014)	**0.015** (0.015)	**0.004** (0.004)
deBias-OLS	0.064 (0.047)	**0.024** (0.006)	0.040 (0.044)	0.010 (0.021)
OLS	**0.016** (0.030)	0.058 (0.015)	0.040 (0.017)	**0.004** (0.004)
Bayesian (uniform)	0.459 (0.014)	0.357 (0.138)	0.464 (0.014)	0.418 (0.073)
Bayesian (Guassian)	0.285 (0.045)	0.283 (0.070)	0.337 (0.020)	0.290 (0.047)
L_2 Calibration	0.107 (0.171)	0.045 (0.021)	0.062 (0.087)	0.033 (0.061)
Neural Networks I	0.256 (0.158)	0.056 (0.017)	0.121 (0.148)	0.078 (0.093)
Neural Networks II	0.130 (0.232)	0.052 (0.018)	0.020 (0.021)	0.023 (0.047)

5. Conclusion

This study presents a novel bias-corrected parameter calibration approach to effectively calibrate the BEM, while simultaneously mitigating the heterogeneous variance of the electricity consumption data. This approach enables us to design a new algorithm that explores the IRLS method in linear regression. Specifically, we analyze the systematic bias between the actual and simulated electrical energy consumption pattern present in the BEM. We show that this pattern can be captured by using the time-series model that incorporates seasonal components. Moreover, the variance of residuals may exhibit heterogeneous patterns, especially the inflated variance in the afternoon. This is often well-explained by the stochastic and heterogeneous occupant behavior in the residential building (e.g., starting to turn on the air conditioner, etc.) This heterogeneity may negatively affect prediction and uncertainty quantification capabilities. To address this heterogeneity, we introduce weights in the loss function. This procedure can be achieved using the IRLS procedure.

Our implementation results demonstrate that the proposed approach can improve prediction accuracy significantly in terms of several metrics. In particular, both CVRMSE and NMBE results satisfy the industry guidelines in the BEM

calibration. Moreover, we demonstrate the improved uncertainty quantification capabilities through the proposed deBias-IRLS method.

In the future, we aim to broaden the scope of our methodology to encompass more generalized settings. For example, we will consider the outputs from multiple channels. Smart meters allow utility providers to collect electricity consumption data from different channels such as HVAC, lighting, and appliances. Thus, we can calibrate parameters with multi-output data if such data become available to us. Furthermore, we plan to apply the well-calibrated BEM for the purpose of control and management in building energy use as part of demand response programs (Jang et al. 2020, Li et al. 2020), as well as renewable generation planning in microgrid operations (Wang et al. 2021).

Acknowledgement

This work was supported in part by the U.S. National Science Foundation under Grant CMMI-2226348.

References

M. Abramowitz and I. A. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards, U.S. Department of Commerce, 1972.

H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

ASHRAE. Guideline 14-2002, measurement of energy and demand savings. Technical report, American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), 2002.

ASHRAE. Guideline 14-2014, measurement of energy and demand savings. Technical report, U.S. Department of Energy, 2014.

S. Bhatnagar, W. Chang, S. Kim, and J. Wang. Computer model calibration with time series data using deep learning and quantile regression. SIAM/ASA Journal on Uncertainty Quantification, 10(1):1–26, 2022.

A. Booth, R. Choudhary, and D. Spiegelhalter. A hierarchical Bayesian framework for calibrating micro-level models with macro-level data. Journal of Building Performance Simulation, 6(4):293–318, 2013.

E. Byon, Y. Choe, and N. Yampikulsakul. Adaptive learning in time-variant processes with application to wind power systems. IEEE Transactions on Automation Science and Engineering, 13(2):997–1007, 2016.

G. Casella and R. L. Berger. Statistical Inference. Thomson Learning, 2nd edition, 2002.

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

Y. Choe, H. Lam, and E. Byon. Uncertainty quantification of stochastic simulation for black-box computer experiments. Methodology and Computing in Applied Probability, 20(4):1155–1172, 2018.

A. Chong and K. Menberg. Guidelines for the Bayesian calibration of building energy models. Energy and Buildings, 174:527–547, 2018.

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

A. Chong, Y. Gu, and H. Jia. Calibrating building energy simulation models: A review of the basics to guide future work. Energy and Buildings, 253:111533, 2021. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2021.111533.

C. Christensen, S. Horowitz, and U.S. Department of Energy Office of Energy Efficiency and Renewable Energy. Beopt™ (building energy optimization tool) [swr-05-41], version 2.8.0.0, 11 2011.

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

P. De Wilde. The gap between predicted and measured energy performance of buildings: A framework for investigation. Automation in Construction, 41:40–49, 2014.

S. De Wit and G. Augenbroe. Analysis of uncertainty in building design evaluations and its implications. Energy and Buildings, 34(9):951–958, 2002.

J. J. Faraway. Linear Models with R. Chapman and Hall/CRC, 2nd edition, 2014.

P. Frazier. A tutorial on Bayesian optimization. ArXiv, abs/1807.02811, 2018.

P. Frazier, W. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal beliefs. INFORMS Journal on Computing, 21(4): 599–613, 2009.

J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes. Accuracy of automated measurement and verification (m&v) techniques for energy savings in commercial buildings. Applied Energy, 173:296–308, 2016.

T. Hastie, R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, volume 2. Springer, 2009.

Y. Heo, D. J. Graziano, L. Guzowski, and R. T. Muehleisen. Evaluation of calibration efficacy under different levels of uncertainty. Journal of Building Performance Simulation, 8(3):135–144, 2015.

D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing, 26(2):448–466, 2004.

S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, nov 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

P. Jain, S. Shashaani, and E. Byon. Wake effect parameter calibration with large-scale field operational data using stochastic optimization. Applied Energy, 347:121426, 2023.

Y. Jang, E. Byon, E. Jahani, and K. Cetin. On the long-term density prediction of peak electricity load with demand side management in buildings. Energy and Buildings, 228:110450, 2020.
Y. Jang, E. Byon, S. Vanage, K. Cetin, D. E. Jahn, W. Gallus, and L. Manuel. Spatiotemporal post-calibration in a numerical weather prediction model for quantifying building energy consumption. IEEE Transactions on Automation Science and Engineering, 20(4):2732–2747, 2023.
C. Jeong, X. Ziang, E. Byon, A. S. Berahas, and K. Cetin. Multi-block parameter calibration in computer models. INFORMS Journal on Data Science, 1:160, 2023.
D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4): 455–492, 1998.
M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63:425–464, 2001.
Y. J. Kim and C. S. Park. Stepwise deterministic and stochastic calibration of an energy simulation model for an existing building. Energy and Buildings, 133:455–468, 2016.
Y.-S. Kim, M. Heidarinejad, M. Dahlhausen, and J. Srebric. Building energy model calibration with schedules derived from electricity use data. Applied Energy, 190:997–1007, 2017.
M. H. Kristensen, R. Choudhary, and S. Petersen. Bayesian calibration of building energy models: comparison of predictive accuracy using metered utility data of different temporal resolution. Energy Procedia, 122:277–282, 2017.
M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. Applied Linear Statistical Models. McGraw-Hill, 5th edition, 2005.
G. Lee, E. Byon, L. Ntaimo, and Y. Ding. Bayesian spline method for assessing extreme loads on wind turbines. Annals of Applied Statistics, 7 (4):2034–2061, 2013.
Y.-S. Lee and L.-I. Tong. Forecasting nonlinear time series of energy consumption using a hybrid dynamic model. Applied Energy, 94:251–256, 2012.
D. Li, C. C. Menassa, V. R. Kamat, and E. Byon. Heat-human embodied autonomous thermostat. Building and Environment, 178:106879, 2020.
Q. Li, G. Augenbroe, and J. Brown. 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, 2016.
B. Liu, X. Yue, E. Byon, and R. Kontar. Parameter calibration in wake effect simulation model with stochastic gradient descent and stratified sampling. Annals of Applied Statistics, 2021.
J. L. Loeppky, J. Sacks, and W. J. Welch. Choosing the sample size of a computer experiment: A practical guide. Technometrics, 51(4):366–376, 2009.
X. Lü, T. Lu, C. J. Kibert, and M. Viljanen. Modeling and forecasting energy consumption for heterogeneous buildings using a physical–statistical approach. Applied Energy, 144:261–275, 2015.
M. Manfren, N. Aste, and R. Moshksar. Calibration and uncertainty analysis for computer models: a meta-model based approach for integrated building energy simulation. Applied Energy, 103:627–641, 2013.
E. Mantesi, C. J. Hopfe, M. J. Cook, J. Glass, and P. Strachan. The modelling gap: Quantifying the discrepancy in the representation of thermal mass in building simulation. Building and Environment, 131:74–98, 2018.
M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2):239–245, 1979.
K. Menberg, Y. Heo, and R. Choudhary. Efficiency and reliability of Bayesian calibration of energy supply system models. Proceedings of the 15th IBPSA Building Simulation Conference, 2017.
A. C. Menezes, A. Cripps, D. Bouchlaghem, and R. Buswell. Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap. Applied Energy, 97:355–364, 2012.
J. Močkus. On Bayesian methods for seeking the extremum. In Optimization techniques IFIP technical conference, pages 400–404. Springer, 1975.
M. D. Morris and T. J. Mitchell. Exploratory designs for computational experiments. Journal of Statistical Planning and Inference, 43(3):381–402, 1995. ISSN 0378-3758. doi: https://doi.org/10.1016/0378-3758(94)00035-T.
Q. Pan, Y. M. Ko, and E. Byon. Uncertainty quantification for extreme quantile estimation with stochastic computer models. IEEE Transactions on Reliability, 70(1):134–145, 2021.
R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.
C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning. Adaptive computation and machine learning. MIT Press, 2006.
O. Roustant, D. Ginsbourger, and Y. Deville. Dicekriging, diceoptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software, 51(1):1–55, 2012.
T. J. Santner, B. J. Williams, and W. I. Notz. The Design and Analysis of Computer Experiments. Springer, 2nd edition, 2018.
G. Schwarz. Estimating the dimension of a model. Annals of Statistics, 6(2):461–464, 1978.
B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1):148–175, 2015.
R. H. Shumway and D. S. Stoffer. Time Series Analysis and Its Applications with R Examples. Springer, 4th edition, 2017.
J. Sokol, C. C. Davila, and C. F. Reinhart. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy and Buildings, 134:11–24, 2017.
N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. Proceedings of the 27th International Conference on International Conference on Machine Learning, page 1015–1022, 2010.
W. Tian, S. Yang, Z. Li, S. Wei, W. Pan, and Y. Liu. Identifying informative energy data in Bayesian calibration of building energy models. Energy and Buildings, 119:363–376, 2016.
R. Tuo and J. C. Wu. Efficient calibration for imperfect computer models. Annals of Statistics, 43:2331–2352, 2015.
C. Turner, M. Frankel, and U. Council. Energy performance of leed for new construction buildings. New Buildings Institute, 4(4):1–42, 2008.
U.S. Department of Energy. Eneryplus essentials. Technical Report, 2019.
U.S. Energy Information Administration. Table 2.1 Energy Consumption by Sectors, 2015.

J. Wang, S. Chung, A. AlShelahi, R. Kontar, E. Byon, and R. Saigal. Look-ahead decision making for renewable energy: A dynamic "predict and store" approach. Applied Energy, 296:117068, 2021.

L. Webster and J. Bradford. M & v guidelines: Measurement and verification for federal energy projects. Technical report, U.S. Department of Energy Office of Energy Efficiency and Renewable Energy, 2002.

L. Webster, J. Bradford, D. Sartor, J. Shonder, E. Atkin, S. Dunnivant, D. Frank, E. Franconi, D. Jump, S. Schiller, M. Stetz, and B. Slattery. M & v guidelines: Measurement and verification for federal energy projects. Technical report, U.S. Department of Energy Office of Energy Efficiency and Renewable Energy, 2014.

Z. Xu, C. Jeong, E. Byon, and K. Cetin. Season-dependent parameter calibration in building energy simulation. Proceedings of 2021 IISE Annual Conference, pages 423–428, 2021.

H. Yoshino, T. Hong, and N. Nord. Iea ebc annex 53: Total energy use in buildings—analysis and evaluation methods. Energy and Buildings, 152: 124–136, 2017.