Retraining surrogate models in increasingly restricted design spaces: A novel building energy model calibration method

Florent Herbinger*[a] (ORCiD: 0000-0003-1715-6709)

Michaël Kummert[a] (ORCiD: 0000-0003-1606-8344)

*Corresponding author: florent.herbinger@polymtl.ca

[a]Affiliation: Polytechnique Montréal, QC, CANADA

Abstract

Surrogate (*i.e.*, meta) models can approximate building energy models (BEMs) accurately and quickly, hence they have been widely used in BEM calibration studies. Typically, the surrogate models are trained a single time over the entire unknown building parameter space with a design such as Latin hypercube sampling. In this article, a multiple polynomial regression surrogate model is, instead, retrained with increasingly restricted designs. In each training repetition, the bounds of the design narrow around the unknown building parameter values that minimize the error between the surrogate model's predictions and the measured energy. This "cascading surrogate" calibration method finds CVRMSE values that are much lower than those of a powerful black box optimizer in a case study with simulated "measured" data. However, the method has similar performance to the black box optimizer in a case study with real hourly measured energy, probably since the BEM was not configured accurately enough.

Keywords

Building energy model; Calibration; Surrogate; Metamodel; Polynomial Regression; Design of experiment.

The Version of Record of this manuscript has been published and is available in Journal of Building Performance Simulation, 29 April 2024, http://www.tandfonline.com/10.1080/19401493.2024.2346833

1 Introduction

With the advent of powerful computers and advances in software, building performance simulation is seeing increased use among building practitioners. Several organizations, including the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE 1975), have long recognized the benefits of having an accurate building energy model (BEM) of a future or existing building; for example, the reduction in energy use from potential retrofits can be estimated, and the building's energy and thermal performance can be assessed to ensure it is working properly and efficiently (*i.e.*, continuous commissioning). However, for building energy simulation to be most effective, the BEM must be *calibrated*. In other words, its output must closely match the historical behaviour of the building being modeled. This way, the BEM closely emulates the actual building's performance, and the impacts from retrofits can be accurately estimated for example. Yet, buildings are highly complex and often have many unknown parameters, leading to the *curse of dimensionality* (Bellman 1957). As a result, even after calibrating a building energy model, there can still be discrepancies of up to 250% between simulated and measured data (de Wilde 2014). New calibration methods should therefore be developed that increase the accuracy of building energy models, not only for aggregated consumption profiles, such as monthly electricity consumptions, but for dynamic hourly and sub-hourly profiles. The electrification of energy grids across the world makes this objective a pressing one.

In this article, we propose an effective calibration method for dynamic profiles based on surrogate modelling and cascading training data distributions. All calibration methods, if they do not have a fundamental flaw in their mathematical formulation, will eventually find the optimal solution given enough optimization time. What differentiates calibration methods is how effective they are at finding a "good enough" solution given a reasonable time constraint. We demonstrate in this article that our calibration method is more effective than the very well-established method of black box optimization.

1.1 Building energy model calibration

Before calibrating a BEM, it is first defined in software like EnergyPlus by fixing the building parameters that are known or safely assumed, such as the building geometry. Then, the remaining unknown building parameters are adjusted, either by a human or a computer, until the simulated output of the BEM (such as the energy consumption) matches the historical measured behaviour of the building. There are two common criteria that are used to assess the calibration level of a BEM: the coefficient of variation of the root mean square error (CVRMSE) and the normalized mean bias error (NMBE). Equation (1) and (2) below show the formula for calculating the CVRMSE and NMBE, respectively:

$$CVRMSE = \frac{1}{\bar{m}} \sqrt{\frac{\sum_{i=0}^{n}(m_i - s_i)^2}{n-p}} \times 100\ (\%) \tag{1}$$

$$NMBE = \frac{1}{\bar{m}} \frac{\sum_{i=0}^{n}(m_i - s_i)}{n-p} \times 100\ (\%) \tag{2}$$

where $\boldsymbol{m_i}$ is the measured value, $\boldsymbol{s_i}$ is the simulated value, $\boldsymbol{\bar{m}}$ is the mean of all measured values, $\boldsymbol{n}$ is the number of measured values, and a recommended value for $\boldsymbol{p}$ is 1 (Reddy et al. 2006). Table 1 shows the common thresholds that are used for these criteria. A BEM with a CVRMSE and NMBE below these thresholds is said to be sufficiently calibrated.

Table 1 : Calibration criteria thresholds for building calibration from various organizations.

Calibration Criteria	**ASHRAE Guideline 14** (ASHRAE 2014)	**FEMP** (U.S. Department of Energy 2015)	**IPMVP** (Cowan 2002)
Monthly CVRMSE	15 %	15 %	-
Hourly CVRMSE	30 %	30 %	20 %
Monthly NMBE	±5 %	±5 %	-
Hourly NMBE	±10 %	±10 %	±5 %

To achieve these thresholds, manual BEM calibration by a modeler is still very common. However, it is generally time-consuming, expensive, and dependent on the modeler's expertise. Automated BEM calibration addresses these concerns by using mathematical and statistical techniques to automatically calibrate the unknown building parameters. The two most common automated calibration methods are deterministic black box optimization and Bayesian optimization (Chong, Gu, and Jia 2021).

Deterministic black box optimization couples a black box algorithm (such as a genetic algorithm) to a building simulation software to let it automatically find the unknown building parameter values that minimize, for example, the CVRMSE or NMBE. During each iteration of the optimization, the algorithm chooses a set of unknown building parameters, and the BEM is simulated with these parameter values. The CVRMSE or NMBE is then calculated between the simulated output of the BEM and the corresponding measured data. At the end of its optimization, the black box algorithm returns the set of unknown building parameter values that resulted in the lowest CVRMSE or NMBE. This calibration method has been widely used in the literature: for example in (Monetti et al. 2015; W. Li et al. 2018; Yang et al. 2016) to only name a few.

Bayesian optimization uses principles from Bayes's theorem to generate a probability distribution over parameter values as opposed to a single estimate, thereby considering the inherent uncertainty in BEM calibration. This method requires input from the building modeler in the form of a prior probability distribution for the unknown building parameter values; in other words, the building modeler must first provide the most probable ranges (and, possibly, values) for the building parameters that explain the measured energy data. Through the Bayesian calibration optimization, this prior probability distribution is then updated to become more and more centered on the most likely values that minimize the aggregated error between simulated and measured energy data. From the final posterior probability distributions, the building energy modeler manually selects the calibrated unknown building parameters that they think are most logical.

In black box optimization and Bayesian optimization, the BEM must be simulated thousands to tens of thousands of times to give the algorithms the best chance to calibrate the unknown building parameters (Gilks, Richardson, and Spiegelhalter 1996). However, with building energy models taking on the order of minutes to simulate, the calibration time can quickly balloon. Surrogate models can greatly reduce the computation time by replacing the BEM. Surrogate models (or metamodels) emulate more complex models, such as a BEM, and can be used to quickly approximate their output. The surrogate model is trained by feeding it many inputs (the unknown building parameters) and the associated output from the complex model (the BEM's energy consumption). It then learns how to map the inputs to the outputs, which allows it to approximate the output of the complex model given new inputs. Several building energy model calibration studies have trained surrogate models of varying types to replace a BEM; subsequently, an algorithm finds the unknown building parameter values or distribution of values that minimize the error between the surrogate model's predictions and the measured energy consumption: (Nagpal et al. 2019; Chen et al. 2019; Yuan et al. 2017; Cant and Evins 2022).

1.2 Comparing surrogate models for building energy model calibration

In the (2021) review paper by Chong, Gu, and Jia, 107 building energy model calibration articles (between the period of 2015 and 2020) were reviewed and classified. They found that 66% of articles employed an automated method rather than a manual method, a three-fold increase compared to Coakley, Raftery, and Keane's findings in their (2014) review paper. Among the automated calibration articles, 58.5% employed deterministic black box optimization and 33% employed Bayesian optimization. However, the exact numbers in this review article should be interpreted with caution as we found errors when digging into the raw data that the authors provide a link to. For example, a surrogate regression model tag was attributed to an article (Ferrara et al. 2020) that clearly did not use a surrogate regression model. Despite these inconsistencies, we still get a general overview of the recent trends in building energy model calibration using surrogate models. The most common type of surrogate model used in the reviewed calibration studies is multiple linear regression (MLR) followed by Gaussian process regression (GPR)—used almost exclusively in Bayesian optimization studies. Random forests (RF), artificial neural networks (ANN), and support vector regression (SVR) are less common with multivariate adaptive regression splines (MARS) being the least common.

In their (2018) article, Østergård, Jensen, and Maagaard compared the performance and efficiency of these aforementioned surrogate models. They studied how well these six surrogate models could predict a single value from a year-long simulation in the BSim software, such as the yearly aggregated energy consumption or the maximum hourly CO_2 level in the building. With the more difficult task of predicting the maximum hourly CO_2 level, most models performed similarly when trained on 2048 or more BEM simulations, apart from the MLR, which performed the worst across all training data sizes. In terms of training time, the MLR surrogate model was by far the quickest and the MARS, GPR and SVR models were the slowest. For the prediction time, the MLR was also the fastest, followed by the RF and SVR models, while the GPR model was the slowest. The ANN model was relatively quick considering how accurate it is. However, as with all other surrogate models in Østergård, Jensen, and Maagaard's study (2018), the ANN model was quite tiny with only 20 hidden neurons in a single layer. The hyperparameters of all six surrogate models were not tuned over a very large range, favouring small models, but considering that only a single value is being predicted by the surrogate models, small models were often sufficient to achieve high prediction accuracy.

There are a few comparative studies where surrogates were used to predict a lot more than a single value, notably the hourly energy consumption of a building. Li et al. (2023) studied a large range of surrogate models for predicting simulated hourly

heating and cooling energy consumption, including gradient boosting machines (GBM), cubist regression (CBR), deep multilayer perceptron ANNs (DNN), and long short-term memory ANNs (LSTM) (in addition to the six surrogate model types previously mentioned). Li et al. studied the effect of including previous independent and dependent variables for prediction in what they call “lag free”, “distributed lag” and “autoregressive lag” surrogate models. The key take-aways for us from this large study is that (1) the tree-based GBM and CBR models performed well, while remaining computationally efficient, and (2) the DNN performed the best while being the most computationally expensive. However, again, we are led to question the choice for the model sizes since a mere 20 neurons in a single layer were used for their ANN surrogates, for example.

1.3 Implementations of surrogate models in building energy model calibration

Based on our literature review, surrogate models have exclusively been trained a single time over the entire unknown building parameter space. They then replace the building energy model when either calibrating through black box optimization (Nagpal et al. 2019; Chen et al. 2019) or Bayesian optimization (Yuan et al. 2017; Cant and Evins 2022). Once trained, the surrogate model replaces the BEM when evaluating the objective function of the black box optimizer or drawing the posterior probability distributions in Bayesian optimization. Since the posterior probability distribution is often drawn with Markov chain Monte Carlo (MCMC), thousands to tens of thousands of BEM simulations are required (Gilks, Richardson, and Spiegelhalter 1996), making the use of fast surrogate models very appropriate. We found that the use of surrogate models is a lot more common in Bayesian optimization than in deterministic black box optimization.

For black box optimization, we only found two studies that used a surrogate model to replace the physics-based BEM. Nagpal et al. (2019) trained 27 different random forests (RFs) and artificial neural network (ANN) surrogate models to predict 12 months of electricity, heating and cooling consumption. The surrogate models were trained a single time over the entire unknown building parameter space with a simple random Latin hypercube sampling design of 400 BEM simulations. The measured electricity consumption was very closely predicted (CVRMSE between 0 and 12%), but the heating and cooling CVRMSEs remained high, ranging from 6 to 62% between the case study buildings and surrogate models. Chen et al. (2019) also trained surrogate models for black box calibration a single time over the entire unknown building parameter space. Multiple linear regression (MLR) and gaussian process regression (GPR) surrogate models were trained to predict 8 monthly electricity consumptions. The authors found that the GPR models outperformed the MLR models, but took longer to train, a finding that is consistent with other studies. Only 50 BEM simulations were required to generate enough training for the GPR model. Like Nagpal et al.’s study (2019), the training data was sampled with a simple random LHS design. The final monthly CVRMSE error in Chen et al.’s case study was ~14%, just below ASHRAE’s guideline of 15%.

In Bayesian optimization studies, the surrogate model is also trained a single time. Yuan et al. (2017) trained a Gaussian process regression (GPR) surrogate model with a simple random LHS of 200 BEM simulations. The GPR model was trained to predict 18 monthly electricity consumptions, and the final CVRMSE was 7.7% for the training data and 8.6% on the 6-month validation data. Cant and Evins (2022) trained a two-layer perceptron neural network with a simple random LHS of 300 BEM simulations. The neural network predicted two monthly electricity profiles and one monthly gas profile between 2017 and 2019, for a total of 108 points, achieving CVRMSEs of between 5 and 12 % for the maximum a-posterior (MAP) parameter values, which are the values with the highest likelihood following the Bayesian calibration.

2 Methodology

2.1 Overview of proposed calibration method

When surrogate models are used in BEM calibration studies, they are trained a single time with a design of experiments (DoE) over the entire unknown building parameter space. In this article, a surrogate is trained several times over an increasingly restricted design space, much like water flowing over a narrowing cascade. The design space is restricted by narrowing the bounds of the training data around the unknown building parameters that minimize the error between the surrogate model’s predictions and the measured energy. Surrogate models that are trained on the same amount of data in a smaller range have a smaller distance over which to interpolate, resulting in more accurate predictions in that space. By retraining the MPR surrogate several times in a design space that appears to get closer and closer to the global minimum, the MPR surrogate’s estimates become more and more accurate, achieving a positive feedback loop towards the apparent global minimum.

This cascading calibration method is fundamentally different than black box optimization, since it does not just narrow down the "area of search" like a gradient descent optimizer does, but it also concentrates the collection of sample data towards a smaller and smaller region of the parameter space. It concentrates the sample data in the region of the current solution so that a very locally accurate surrogate model can be obtained.

To ensure that the sampled training data zeros-in on a good local minimum, it is imperative that the surrogate model emulates the BEM as accurately as possible; however, the training time of the surrogate model and the estimation time of the unknown building parameters should still be reasonable. We developed a surrogate model that is both accurate and fast, since it fits a lightweight nonlinear function to each hour of the calibration period separately. By fitting each hour separately, the surrogate model is more flexible and can account for how the interior and exterior conditions (*i.e.*, weather) of the building change every hour. The surrogate model used in this cascading surrogate calibration method is a multiple polynomial regression (MPR) model, made up of 8760 different ridge regressions—one for every hour of the year. Ridge regression is ordinary least squares (OLS) regression with added regularization in the form of a penalty on the magnitude of the regression coefficients. This regularization helps with overfitting with overfitting by penalizing large coefficients, thereby discouraging the model from fitting the noise in the training data. Ridge regression also helps when there is collinearity between predictors. In OLS regression with collinear predictors, small changes in the training data can wildly swing the estimated coefficients. By penalizing the coefficients towards zero, small changes do not alter the estimated coefficients nearly as much.

The cascading surrogate calibration method follows four-steps (see Figure 1)

Step 1: Use orthogonal-array-based Latin hypercube sampling (LHS) to generate different sets of unknown building parameter values (X_{params}) with wide initial bounds. Run the physics-based BEM with these sampled points and record the associated hourly simulated heating and cooling consumption (Y_{sim}).

Step 2: Train the surrogate model in a supervised way by regressing the unknown inputs (X_{params}) onto the BEM's simulated energy consumption (Y_{sim}) for each hour of the year.

Step 3: Perform gradient descent over the surrogate model's parameters to estimate the unknown building parameters (X_{params}') that minimize the difference between the surrogate model's prediction (Y_{surr}) and the measured energy data (Y_{meas}) over the entire year.

Step 4: Generate a new orthogonal-array-based LHS design with unknown building parameter values that have been narrowed around the estimates of the MPR model. Next, run the BEM with these new points and repeat steps 2 to 4 a predefined number of times, returning the calibrated unknown building parameter values at the end.

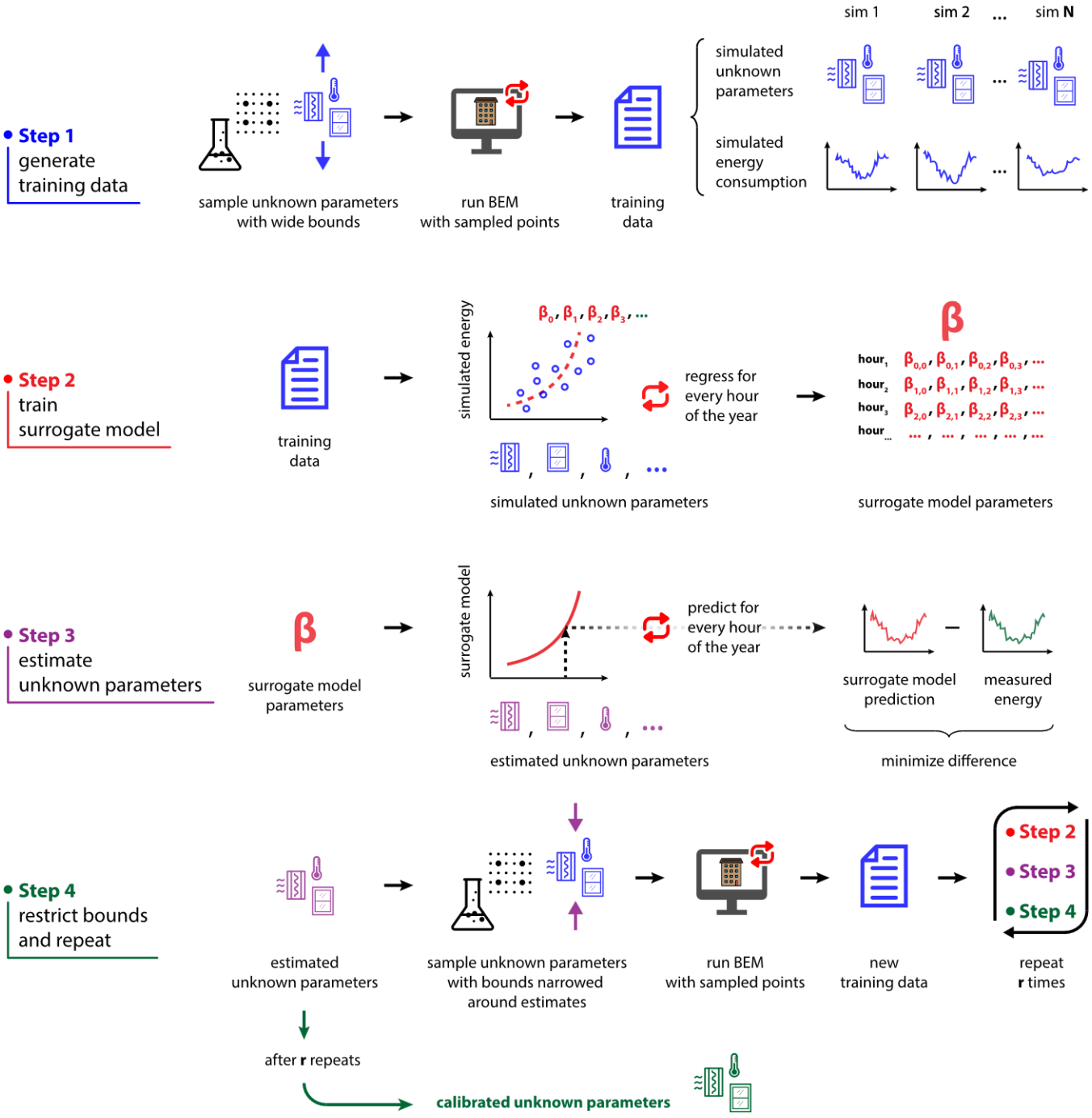

Figure 1: Four-step cascading surrogate calibration method

2.2 Definition of the Surrogate Model

Different surrogate models can be used to emulate a BEM. This section presents the novel multiple polynomial regression (MPR) surrogate model that we developed. The following sections describe the four-step calibration process that uses the surrogate model to calibrate the values of the unknown building parameters.

We first tried a multiple linear regression surrogate model as a baseline. The predictors used in this linear model are simply the unknown building parameters of the BEM, presented in the case study section of the article (Table 4). These predictors were divided into two groups: one for predicting the heating energy consumption and one for the cooling, as shown in Table 2.

Table 2: Unknown building parameter predictors

Heating predictors	Cooling predictors
$\dot{v}_{air,new}$	$\dot{v}_{air,new}$
c_{inf}	c_{inf}
$ctrl_{eff}$	$ctrl_{eff}$
U_{ins}	U_{ins}
$T_{heat,day}$	$T_{cool,day}$
$\Delta T_{heat,setback}$	$\Delta T_{cool,setup}$
ρ_{occ}	ρ_{occ}
$f_{elec2heat}$	$f_{elec2heat}$
$SHGC_w$	$SHGC_w$
$\dot{v}_{shw}$	$loss_{cool}$
$loss_{heat}$	

This linear surrogate model performed relatively poorly in a controlled case study (where the "calibrated" values of the unknown building parameters are known ahead of time). Notably, the unknown setpoint temperature parameters were poorly estimated. Consequently, the relationship between the unknown building parameters and the heating and cooling consumption was more closely examined by plotting the simulated heating or cooling energy consumption at a given hour against a single unknown building parameter, while holding all other parameters fixed. For each varying parameter, 100 values were evenly sampled between the bounds set in the case study in Table 4.

Figures 2 and 3 display these plots for all the heating and cooling unknown building parameters at three different hours of the year; the hours were selected to show the variety of relationships between the unknown building parameters and the simulated energy consumption.

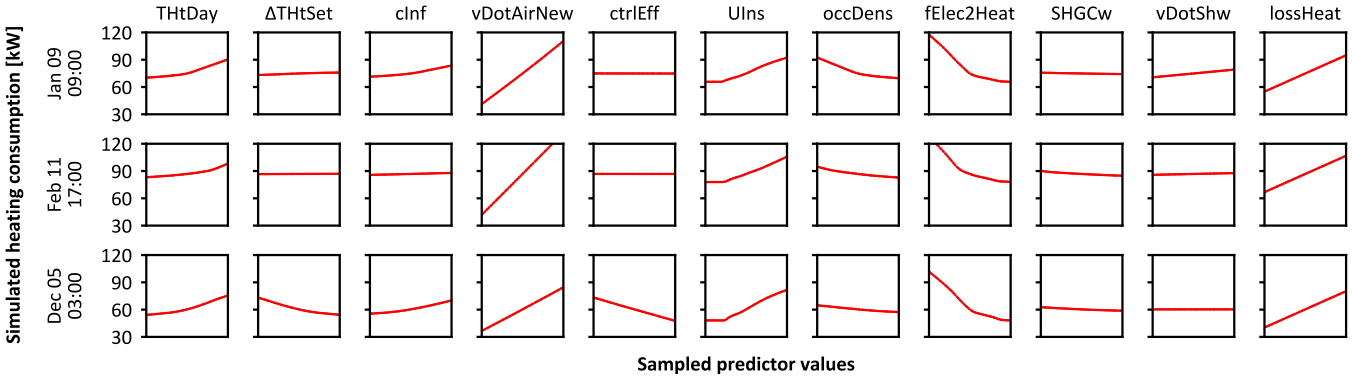

Figure 2: Correlation between individual isolated building parameters and the heating energy consumption for three different hours of the year

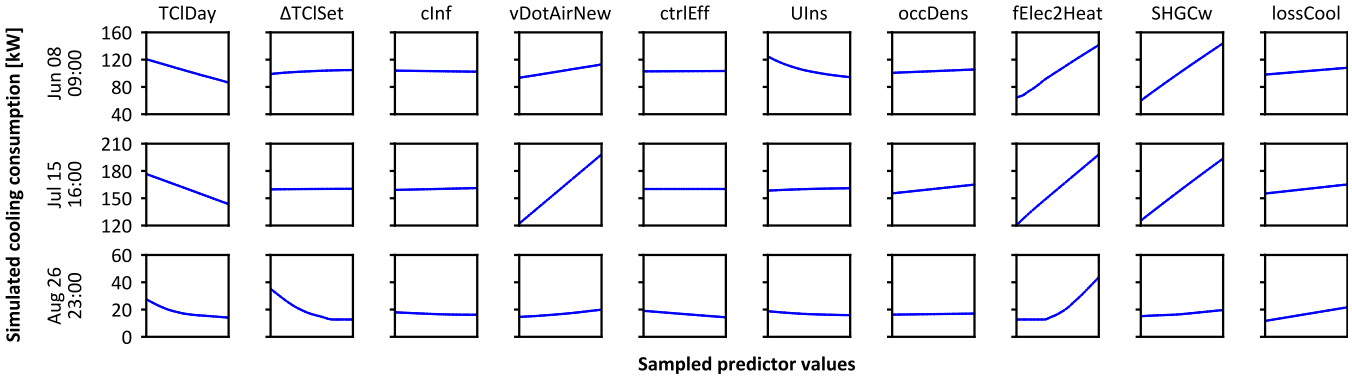

Figure 3: Correlation between individual isolated building parameters and the cooling energy consumption for three different hours of the year

Certain terms, notably the temperature setpoints, the envelope insulation, U_{ins}, and the fraction of electricity that is turned into heat, $f_{elec2heat}$, show marked non-linearity with the energy consumption. This behavior would explain why the temperature setpoint parameters, among others, were poorly estimated in the linear surrogate model. We therefore decided to increase the non-linearity of the surrogate model, with the important consideration that the surrogate model training time and the unknown building parameter estimation time should be as short as possible so that the cascading surrogate calibration method completes in a reasonable time. We tested the following non-linear surrogate models:

- Natural cubic splines,
- P-splines,
- Multivariate adaptive regression splines (MARS),
- Support vector regression (SVR),
- Multiple polynomial regression (MPR).

The training time for all these non-linear surrogate models is acceptable; however, only the multiple polynomial regression model has an unknown building parameter estimation time that is short enough to be incorporated into our calibration method. The reason is that the MPR surrogate model has a prediction function that is much faster to execute in our non-traditional, but more accurate, unknown building parameter estimation technique (see step 3 of the calibration method). However, we could also have used a larger variety of surrogate models if we had decided to use a more classic, black box estimation technique at the expense of some parameter estimation accuracy.

The predictors used in the chosen MPR surrogate model are a polynomial expansion of the unknown building parameters in Table 4 (Equation (3)):

$$h_\delta(X) = X^\delta \qquad \delta = 1, \dots, d, \tag{3}$$

where $\boldsymbol{h_\delta}$ denotes the $\boldsymbol{\delta}$th transformation of an unknown building parameter and $\boldsymbol{d}$ is the degree of the polynomial. $\boldsymbol{d}$ was selected through a grid search (see section 2.4.1).

2.3 Four-step calibration method

In the previous section, we defined the multiple polynomial regression surrogate model. In this section, we describe the four-step BEM calibration method where the MPR surrogate model is trained on increasingly restricted parameter spaces. The calibration method uses parallel CPU threads for the first three of the four steps, while the final single-threaded step is fast and does not require multiprocessing. The calibration method was executed on a system with a 3.4 GHz, 16-core AMD Ryzen 5950X CPU (with 32 threads) and was entirely programmed in a Python package, which the authors hope to make publicly available soon.

2.3.1 Step one: Generate training data

By running multiple simulations of our building energy model with various combinations of values for the unknown parameters, we can generate training data for our surrogate model. It can then map the inputs (the unknown building parameters) to the outputs (the simulated energy consumption) of our building energy model, thereby emulating it.

To generate the data for the surrogate model, a number of BEM simulations need to be executed in the chosen building simulation software (TRNSYS in our case). A design of experiments (DoE) determines which unknown building parameter inputs should be used to generate the BEM's simulated output. When developing surrogate models that predict unsampled points in a space, DoEs that sample the input space evenly and thoroughly (*i.e.*, space-filling) are preferred over designs that are biased towards certain regions of the space (Tang and Lin 2015). Simple random Latin hypercube sampling (LHS) is one of the most common types of DoEs for training surrogate models of BEMs (Westermann and Evins 2019). Random LHS is a multidimensional extension of the *Latin square* in two dimensions, where there is exactly one sample in each row and column. One disadvantage with random LHS is that it may perform poorly at maximizing the distance between points, so points may not be spread very evenly over the design space. Random LHS designs may also not be very orthogonal, meaning that points can line up along diagonals, resulting in collinearity between the inputs. This collinearity reduces the accuracy of surrogate models since they have a harder time distinguishing how collinear inputs affect the output (Tang and Lin 2015).

The DoE chosen in our calibration method is an orthogonal-array-based LHS, developed by Boxin Tang (1993), which is an LHS design of strength 2. A strength 2 LHS DoE considers how the response is affected by each input on its own and by interactions between pairs of inputs. More importantly, this DoE exhibits better space-filling and orthogonal properties than random LHS designs, which are of strength 1 and do not consider two-input interactions. Orthogonality is especially useful for polynomial surrogate models, where inputs are directly correlated by the nature of powers. The number of samples $\boldsymbol{N}$ in an orthogonal-array-based LHS must be the square of a prime number; we selected $\boldsymbol{N}$ through a grid search—see section 0. Once the $\boldsymbol{N}$

BEMs are simulated, 8760 $\boldsymbol{y_h}$ vectors, one for each hour of the year, are compiled for both the simulated heating and cooling consumption (Equation (4)):

$$\boldsymbol{y_h} = \begin{bmatrix} e_{1,h} \\ \vdots \\ e_{n,h} \\ \vdots \\ e_{N,h} \end{bmatrix} \qquad h = 1, \dots, 8760, \tag{4}$$

$\boldsymbol{y_h}$ is the simulated heating or cooling consumption, where $\boldsymbol{e_{n,h}}$ represents the heating or cooling consumption at hour $\boldsymbol{h}$ for building simulation $\boldsymbol{n}$. Additionally, the building parameters are compiled into two $\boldsymbol{X}$ matrices, one for the heating predictors and one for the cooling predictors (Equation (5)):

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} & \cdots & x_{1,P} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} & \cdots & x_{n,P} \\ \vdots & & \vdots & & \vdots \\ x_{N,1} & \cdots & x_{N,p} & \cdots & x_{N,P} \end{bmatrix}, \tag{5}$$

$\boldsymbol{X}$ is composed of vectors of either the heating or cooling predictors, where $x_{n,p}$ is the standardized value of predictor $\boldsymbol{p}$ for building simulation $\boldsymbol{n}$. $\boldsymbol{P}$ is the total number of heating or cooling predictors. The predictors are standardized by subtracting the mean and scaling to unit variance, which can help with training accuracy.

2.3.2 Step two: Train the surrogate model in a supervised way

We now regress both heating and cooling matrices $\boldsymbol{X}$ onto the respective response vector $\boldsymbol{y_h}$. Because there is collinearity between the predictors (*e.g.*, between the occupancy- and electricity-related parameters), ridge regression as opposed to ordinary least squares regression is employed. In ridge regression, a regularization hyperparameter $\boldsymbol{\alpha}$ penalizes the magnitude of the coefficients; the value of $\boldsymbol{\alpha}$ is determined for each individual ridge regression through a cross validation grid search with 50 samples, while the lower and upper bounds of $\boldsymbol{\alpha}$ in the 8760 grid searches is predetermined through a grid search (see section 2.4.1).

$$\min_{\beta_h} \; | X\beta_h - y_h|_2^2 + \alpha \, |\beta_h|_2^2 \, , \tag{6}$$

$\boldsymbol{\beta_h}$ is the vector of the fitted heating or cooling coefficients for hour $\boldsymbol{h}$. We then define $\boldsymbol{\beta}$ as the matrix of concatenated $\boldsymbol{\beta_h}$ vectors for heating or cooling so that $\boldsymbol{\beta}$ has 8760 rows. Early stopping, which is a machine learning technique that stops the training of the model before it overfits the training data, is not necessary with the MPR surrogate model. Our surrogate model is rather simple and small, and it is regularized through ridge regressions, so it is unlikely to overfit the training data.

2.3.3 Step three: Estimate the unknown building parameter values

The $\boldsymbol{\beta}$ coefficients computed in step one are the parameters of our trained surrogate model of the BEM. Step three consists of using the surrogate model parameters and the measured energy consumption to estimate the unknown building parameter values. Typically, what is done in the literature is to treat the surrogate model as a black box and use an external optimizer (such as a genetic algorithm) to find the unknown building parameters that minimizes the difference between the measured energy and the surrogate model's prediction. We have, however, discovered a much faster and more accurate way of estimating the unknown building parameters using a surrogate model. We leverage the fact that we know the trained surrogate model's parameters (it is not in fact a black box) and parametrize the surrogate model's predictions. In other words, we make predictions using the previously estimated $\boldsymbol{\beta_{heat}}$ and $\boldsymbol{\beta_{cool}}$ coefficients and polynomial transformations of the unknown building parameters. We then minimize the squared difference between the measured heating and cooling consumption and the surrogate model's heating and cooling prediction through gradient descent over the 8760 hours of the year:

$$\min_{\omega} \left| \beta_{heat} f_{heat}(\omega) - y_{meas,heat} \right|_2^2 + \left| \beta_{cool} f_{cool}(\omega) - y_{meas,cool} \right|_2^2 \quad s.t \; bnd_{low,orig} \leq \omega \leq bnd_{up,orig}, \tag{7}$$

where $\boldsymbol{f_{heat}(\omega)}$ and $\boldsymbol{f_{cool}(\omega)}$ are the appropriate heating and cooling polynomial transformations of the unknown building parameters $\boldsymbol{\omega}$; $\boldsymbol{y_{meas,heat}}$ and $\boldsymbol{y_{meas,cool}}$ are vectors of length 8760, representing the hourly measured heating and cooling energy; and $\boldsymbol{bnd_{low,orig}}$ and $\boldsymbol{bnd_{up,orig}}$ are the original lower and upper bounds of the unknown building parameters. This method for estimating the unknown building parameters always finds the best solution for the trained surrogate model thanks to the apparent convexity of the problem.

The optimization Python package called Gekko (Beal et al. 2018) is used for this task as it can compile the surrogate model's parameterized predictions for each hour of the year into byte code, meaning that the prediction function of the surrogate

model is not continuously executed in a more costly, high-level environment (like Python) during the gradient descent. This compilation makes the estimation of the unknown building parameters very fast so that it can be done multiple times during the calibration method. Once the Gekko solver has reached convergence, the resulting building parameter estimates are recorded.

2.3.4 Step four: Restrict the bounds of a new DoE and repeat steps two to four.

The recorded unknown building parameter estimates are now used to define a new DoE and generate another training dataset. The DoE is still an orthogonal-array-based Latin hypercube sampling design, but the lower and upper bounds of the unknown building parameters are narrowed around the previously estimated unknown building parameter values. The new bounds for each unknown building parameter are calculated as shown in Equations (8) to (9):

$$bnd_{low,new} = max\left(est - \frac{1}{2}\left(bnd_{up,old} - bnd_{low,old}\right)(1 - f_{reduc}),\quad bnd_{low,orig}\right), \tag{8}$$

$$bnd_{up,new} = min\left(est + \frac{1}{2}\left(bnd_{up,old} - bnd_{low,old}\right)(1 - f_{reduc}),\quad bnd_{up,orig}\right), \tag{9}$$

where $\boldsymbol{bnd_{low,new}}$ and $\boldsymbol{bnd_{up,new}}$ are the lower and upper parameter bounds for the new DoE; $\boldsymbol{bnd_{low,old}}$ and $\boldsymbol{bnd_{up,old}}$ are the lower and upper parameter bounds from the previous DoE; $\boldsymbol{bnd_{low,orig}}$ and $\boldsymbol{bnd_{up,orig}}$ are the original, wide lower and upper parameter bounds from the very first DoE; $\boldsymbol{est}$ are the current unknown building parameter estimates; and $\boldsymbol{f_{reduc}}$ is a parameter bound reducing factor that is determined in a hyperparameter grid search (see section 0). Lastly, if any of the new parameter bounds go beyond the original parameter bounds, we cut them off at the original parameter bounds (through the $\boldsymbol{max}$ and $\boldsymbol{min}$ functions).

Having calculated new parameter bounds, a new orthogonal-array-based LHS design samples a series of $\boldsymbol{N}$ points within these bounds. With this new training data, steps two to four of the cascading surrogate calibration method are repeated a number $\boldsymbol{r}$ times. $\boldsymbol{r}$ is selected through a hyperparameter grid search as described in section 0.

After $\boldsymbol{r}$ repeats, the unknown building parameter estimates at the end of step three become the calibrated unknown building parameter values, and the final combined CVRMSE of the physics-based BEM with these calibrated values is calculated (see Equation (13)).

2.4 Hyperparameter tuning

There are hyperparameters of both the surrogate model and the calibration method as a whole that must be tuned. Hyperparameters are parameters that impact the accuracy of the machine learning model (or a method that uses the model) but that are tuned by testing the model/method instead of training the model. Tuning of hyperparameters is typically achieved with a grid search through different sets of hyperparameter values, with the goal of finding the hyperparameter values with the lowest testing error. Two different grid searches were performed. The first one tests the surrogate model's hyperparameters by training the surrogate model on a training dataset and then calculating the error between the trained surrogate model's predictions and a testing dataset. The second grid search tests the calibration method's hyperparameters by comparing the method's final output (the estimated unknown building parameters) to a testing output. Using separate training and testing datasets in both grid searches ensures that the surrogate model and the calibration method are tested on unknown data points, making them more generalizable.

It should be noted that the tuning of the surrogate model's hyperparameters and the calibration method's hyperparameters only need to be completed once in this building calibration study. It is also possible that the hyperparameter values could be used for calibrating other building energy models. The fact that the surrogate model is a series of simple, regularized regressions helps keep the surrogate model and its hyperparameters generalizable. On the other hand, the calibration method's hyperparameters appear to be dependent on both the method's innerworkings and the noise in the testing dataset. Tuning the hyperparameters using a simulated testing dataset could increase the calibration method's performance in a controlled case study with simulated data, but it could also worsen its performance in a case study with more noisy, real metered data. However, given that these questions have not been fully explored yet, in the following comparative analyses between the cascading calibration method and the black box calibration method, we do not include the hyperparameter tuning time in the total calibration time.

2.4.1 Grid search of the surrogate model's hyperparameters

Three hyperparameters were searched through for the surrogate model: the degree $\boldsymbol{d}$ of the polynomial, the lower bound of regularization parameter $\boldsymbol{\alpha}$, and the upper bound of regularization parameter $\boldsymbol{\alpha}$. We tested 75 different hyperparameter combinations with the possible values for each hyperparameter shown in Table 3.

The inputs of the training dataset are generated by sampling $\boldsymbol{N}$ = 2809 different sets of unknown building parameter values within the original bounds found in Table 4. Orthogonal-array-based Latin hypercube sampling (LHS) is used as the design of experiment (DoE) (hence why $\boldsymbol{N}$ is the square of a prime number). Then to generate the heating and cooling energy outputs, the BEM is simulated 2809 times using these sampled unknown building parameter values.

The testing dataset consists of $\boldsymbol{N}$ = 702 BEM simulations, making the training/testing split 80/20, which is commonly used in machine learning. A simple random LHS DoE samples the 702 unknown building parameter sets within the original bounds set in Table 4. We chose a large number of BEM simulations (*i.e.*, 2809 and 702) to limit the dependency of the grid search on the training and testing dataset distributions.

The testing error for the grid search is calculated as follows (Equation (10)):

$$err_{test} = \sum_{i=1}^{i=N} \frac{1}{N} \left(\left| \beta_{heat} f_{heat}(\omega_{samp,i}) - y_{test,heat,i} \right|_2^2 + \left| \beta_{cool} f_{cool}(\omega_{samp,i}) - y_{test,cool,i} \right|_2^2 \right), \tag{10}$$

where $\boldsymbol{N}$ is the number of BEM simulations in the testing dataset, $\boldsymbol{\omega_{samp,i}}$ are the sampled unknown building parameter values for the $\boldsymbol{i}$th BEM simulation, $\boldsymbol{f_{heat}}$ and $\boldsymbol{f_{cool}}$ are the polynomial functions of the MPR model, $\boldsymbol{\beta_{heat}}$ and $\boldsymbol{\beta_{cool}}$ are the polynomial coefficients of the trained MPR model, and $\boldsymbol{y_{test,heat,i}}$ and $\boldsymbol{y_{test,cool,i}}$ are the yearly heating and cooling energy consumptions of the $\boldsymbol{i}$th BEM simulation in the testing dataset.

The grid search took ~8.4 hours to complete. The best hyperparameter values, shown in Table 3, reveal that cubic ridge regressions with relatively strong regularization values offered the best prediction accuracy. The clear non-linear relationship between the building parameters and energy consumption is accounted for, while avoiding overfitting the training data. That being said, the relative error difference between the best and worst performing hyperparameter set was only ~12%.

Table 3: Hyperparameters of the multiple polynomial regression surrogate model

Hyperparameter description	Hyperparameter name	Possible grid search values	Best value
Degree of the polynomial	$\boldsymbol{d}$	2, 3, 4	3
Regularization parameter range for the cross validation grid search in the ridge regressions	Lower bound of $\boldsymbol{\alpha}$	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}	10^{-2}
	Upper bound of $\boldsymbol{\alpha}$	10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}	10^{0}

2.4.2 Grid search of the calibration method's hyperparameters

Having found the optimal MPR surrogate model hyperparameter values, we proceeded with a grid search of the three hyperparameters of the calibration method as a whole: the parameter bound reducing factor $\boldsymbol{f_{reduc}}$, the number of sampled points in the LHS designs $\boldsymbol{N}$, and the number of LHS design repeats $\boldsymbol{r}$. The optimization was done through a manual grid search this time due to the large number of possible combinations and the time cost of each cascading surrogate calibration run. We first tested a few different hyperparameter values, observed how they impacted the accuracy of the calibration method, and then selected new hyperparameter values to test, repeating as needed.

The testing error is defined differently in this grid search compared to the grid search of the surrogate model's hyperparameters. We are testing the final output of the calibration method: the estimated unknown building parameters. Therefore, we define the testing error as the mean squared error between a testing set of unknown building parameters and the calibration method's estimates (Equation (11)). The testing values were randomly chosen within the bounds set in Table 4.

$$MSE = \frac{1}{N_{prms}} \sum_{i}^{N_{prms}} \left(prm_{scl,test,i} - prm_{scl,est,i} \right)^2. \tag{11}$$

$\boldsymbol{prm_{scl,test,i}}$ and $\boldsymbol{prm_{scl,est,i}}$ are the testing and estimated scaled values of the $\boldsymbol{i}$th unknown building parameter, and $\boldsymbol{N_{prms}}$ is the number of unknown building parameters. The unknown building parameters have different dimensions and magnitudes; for example, the possible day temperature setpoint values are on the order of 10 while the infiltration coefficient is on the order of 0.1. Therefore, in the calculation of the $\boldsymbol{MSE}$, we decided to scale the testing and estimated unknown building parameter values between 0 and 1 based on their original lower and upper bounds (see Table 4) so that all unknown building parameters have roughly equal weight in the objective function (Equation (12)):

$$prm_{scl} = \frac{prm - bnd_{low,orig}}{bnd_{up,orig} - bnd_{low,orig}}, \tag{12}$$

where $\boldsymbol{prm_{scl}}$ and $\boldsymbol{prm}$ are the scaled and unscaled unknown building parameter values and $\mathbf{bnd_{low,orig}}$ and $\mathbf{bnd_{up,orig}}$ are the original lower and upper unknown building parameter bounds.

We tested a total of 36 hyperparameter configurations, which took around 5.6 days to complete. The tested configuration with optimal results is a reducing factor of 0.5, an LHS design size of 841, and 9 LHS design repeats (Figure 4). More LHS design repeats and reducing factors of 0.5 and 0.6 offer the best performance. Higher Reducing factors (*i.e.*, $\boldsymbol{f_{reduc}}$ = 0.7) narrow the unknown building parameter bounds too quickly, especially for high numbers of LHS design repeats. LHS design sizes larger than $\boldsymbol{N}$ = 841 do not show any appreciable performance gains, probably since the additional training data is superfluous.

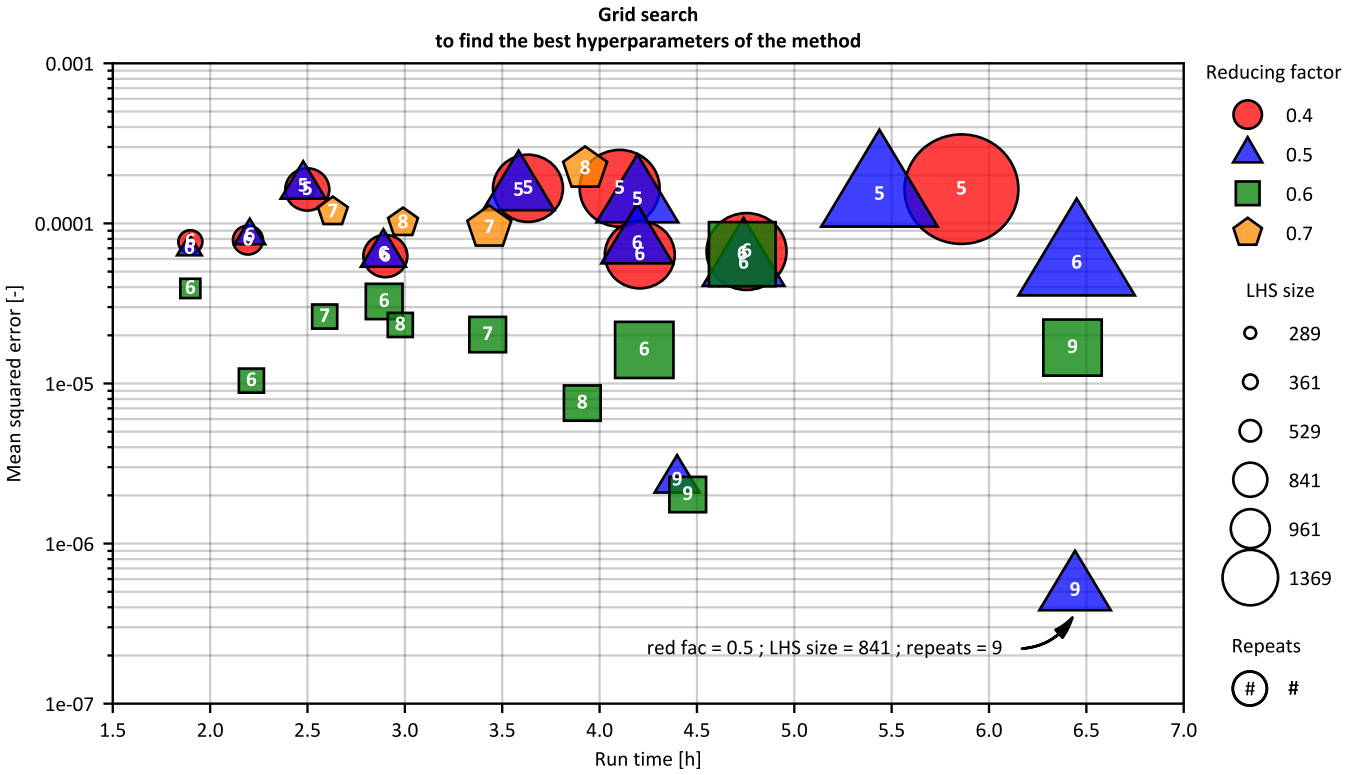

Figure 4: Grid search of the calibration method's hyperparameters of parameter bound reducing factor, Latin Hypercube Sampling (LHS) design size, and number of LHS design repeats.

2.5 Comparing the cascading surrogate calibration method to an established black box optimization method

To test the efficacy of our calibration method, we compared it to a black box optimizer coupled with the detailed BEM. The black box optimization software we selected is RBFOpt from COIN-OR (COmputational INfrastructure for Operations Research), which is a project that provides open-source mathematical software including optimization packages. RBFOpt has proven to be the fastest and most robust open-source algorithm for architectural design as demonstrated by Wortmann's (2019) benchmarking study. Wortmann compared the performance of eight different algorithms across seven simulation-based problems relating to structural, energy, and daylighting optimization, where the number of variables varied from 4 to 40. The types of algorithms tested were evolutionary algorithms (notably genetic algorithms), other direct search algorithms, and model-based algorithms (RBFOpt). The RBFOpt algorithm had the best overall performance among the 8 algorithms tested. RBFOpt is an optimization software optimized for expensive objective function evaluations, making it ideal for BEM calibration.

The objective function that RBFOpt is tasked with minimizing is the average of the heating and cooling CVRMSE, defined in Equation (13). The CVRMSE is the most common criterium (objective function) used in BEM calibration (Chong, Gu, and Jia 2021).

$$\mathrm{CVRMSE}_{\mathrm{comb}} = \frac{1}{2} \times \left(\frac{\sqrt{\frac{\sum_{i=1}^{8760} \left(\mathrm{heat}_{\mathrm{meas,i}} - \mathrm{heat}_{\mathrm{sim,i}}\right)^2}{8760-1}}}{\frac{\sum_{i=1}^{8760} \mathrm{heat}_{\mathrm{meas,i}}}{8760-1}} + \frac{\sqrt{\frac{\sum_{i=1}^{8760} \left(\mathrm{cool}_{\mathrm{meas,i}} - \mathrm{cool}_{\mathrm{sim,i}}\right)^2}{8760-1}}}{\frac{\sum_{i=1}^{8760} \mathrm{cool}_{\mathrm{meas,i}}}{8760-1}} \right), \quad (13)$$

where $\boldsymbol{heat_{meas,i}}$ and $\boldsymbol{heat_{sim,i}}$ are the measured consumption of the building and the simulated energy consumption of the BEM for hour $\boldsymbol{i}$ of the year, and $\boldsymbol{cool_{meas,i}}$ and $\boldsymbol{cool_{sim,i}}$ are the equivalent cooling consumptions. RBFOpt optimizes in 32 parallel threads for a user-defined number of evaluations; during each evaluation, it chooses a set of unknown building parameter values, within the original bounds set in Table 4, and simulates the detailed BEM with these values. It then calculates the average heating and cooling CVRMSE by reading the stored simulated and measured energy files. At the end of the predefined number of evaluations, RBFOpt returns the set of unknown building parameter values that resulted in the lowest combined CVRMSE.

3 Results and Discussion

3.1 Case study building energy model

We used White Hall—an academic building on the Cornell University campus in Ithaca, New York—as our case study building energy model (Figures 5, 6, and 7). This BEM is used to compare the performance of the cascading calibration method with the well-established back box calibration method. The reader is directed to a past paper (Herbinger, Vandenhof, and Kummert 2023) that describes how this building was modeled in the building performance simulation program TRNSYS.

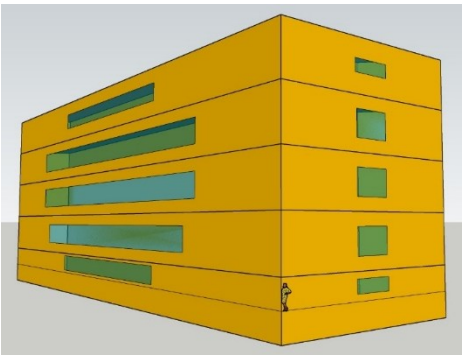

Figure 5: Sketchup model of White Hall with simple, proportional Windows

Figure 6: Google Street view of White Hall

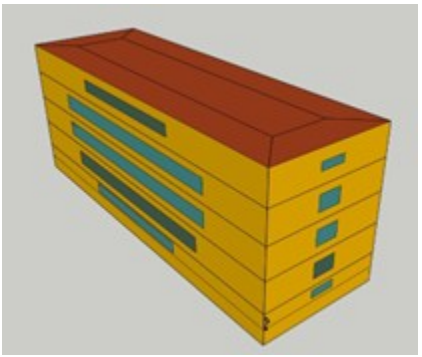

Figure 7: ASHRAE zoning for the White Hall building energy model

3.2 Controlled and real metered case studies

The cascading surrogate calibration method was evaluated in both a controlled and a real-world case study. In the controlled case study, the controlled energy data is in fact energy data generated by a particular configuration of the White Hall BEM. Although the controlled data are necessarily "cleaner" than real metered data, the interest of this case study is that the real values of the parameters of interest are known, and the accuracy of our approach can be assessed.

The real metered case study uses the hourly measured heating and cooling consumption of White Hall during the 2019 year to test the robustness of our method with real data. There is a 20-day period from 00:00 September 27 to 00:00 October 17 where the measured energy consumption is indicative of a fault in the building systems, so this period was not considered in the real metered case study. The measured data therefore covers 345 days in 2019. We direct the reader to our previous paper (Herbinger, Vandenhof, and Kummert 2023) for additional details on the real metered data.

3.3 Unknown building parameters to calibrate

Of all the possible parameters, 14 building parameters were selected to be calibrated, as shown in Table 4. These parameters were selected through a global sensitivity analysis and manual testing as described in our previous article (Herbinger, Vandenhof, and Kummert 2023). The 14 selected parameters were all important for achieving a calibrated model of White Hall. Ranges for the unknown parameters were fixed based on previous modelling experience and domain knowledge for university buildings. Of course, the controlled value of the "unknown" building parameters in the controlled case study had to be within the range.

Table 4: The 14 unknown building parameters to be calibrated.

Category	Building Parameters to be Calibrated	Units	Definition	Range/ Controlled value
Envelope Properties	c_{inf}	s m^{-1} h^{-1}	c infiltration coefficient in $ACH = a + b * deltaT + c * windSpeed + d * windSpeed^2$ (coefficients a, b, and d not considered in this study)	0 – 0.3 Control: 0.1
	U_{ins}	W m^{-2} K^{-1}	Thermal conductance of the massless insulation layer of the exterior walls and roof	0.01 - 3 Control: 0.4
Setpoint temperatures	$T_{heat,day}$	°C	Equivalent average heating setpoint temperature of the building during the day	16 – 20 Control: 19
	$\Delta T_{heat,setback}$	°C	Equivalent average setback of the heating setpoint of the building, which is active during the night. $(T_{heat,night} = T_{heat,day} - \Delta T_{heat,setback})$	0 – 3 Control: 0
	$T_{cool,day}$	°C	Equivalent average cooling setpoint temperature of the building during the day.	20.01 – 24 Control: 23
	$\Delta T_{cool,setup}$	°C	Equivalent average setup of the cooling setpoint of the building, which is active during the night. $(T_{cool,night} = T_{cool,day} + \Delta T_{cool,setup})$	0 – 3 Control: 1
Service Hot Water	$\dot{v}_{shw}$	L min^{-1}	Rated service hot water flow rate. This value is multiplied by a schedule fraction between 0 and 1.	0 – 3 Control: 2
Occupancy	ρ_{occ}	person m^{-2}	Rated occupant density. This value is multiplied by a schedule fraction between 0 and 1.	0 – 0.3 Control: 0.05
Equipment and Lighting	$f_{elec2heat}$	-	Fraction of electricity use that is converted into thermal gains in the building (between 0 and 1)	0.01 – 0.9 Control: 0.7
Solar	$SHGC_w$	-	Global solar heat gain coefficient of the windows	0.05 – 0.7 Control: 0.25
HVAC	$\dot{v}_{air,new}$	$L\ s^{-1}\ m^{-2}$	Fresh air flow rate per unit floor area	0.2 – 1.5 Control: 0.45
	$ctrl_{eff}$	-	The efficacy fraction of the control of the HVAC system. A value of 1 means that the system can shut off completely during the setback period, while lower fractions mean the system is always partially or fully on (for a value of 0)	0 – 0.7 Control: 0.3
	$loss_{heat}$	kW	The baseline heating energy consumption level throughout the year, due to heating equipment losses	0 – 40 Control: 5
	$loss_{cool}$	kW	The baseline cooling energy consumption level throughout the year, due to cooling equipment losses	0 – 10 Control: 10

3.4 Comparing computation times

In Figure 8, we compare the computation times and final combined CVRMSEs between the cascading surrogate method and the RBFOpt black box optimizer method. We show the averaged results from 5 trials for each calibration method to account for some of the inherent randomness in BEM calibration.

With the hyperparameter values optimized for the lowest calibration error, the cascading surrogate calibration method takes around 6.3 hours to calibrate the unknown building parameters. To compare the calibration methods fairly, we fixed the number of RBFOpt optimizer evaluations at 9000 so that the total computation time is also around 6.3 hours.

In Figure 8, we divided up the total computation time into 4 different categories to better illustrate how the calibration methods differ: BEM simulation time, RBFOpt optimization time, surrogate model training time, and unknown building parameter estimation time. We can see that for both methods, simulating the detailed BEM is the longest step. However, the cascading surrogate optimization time—which is the sum of the surrogate model training and unknown building parameter estimation times—is longer than the RBFOpt optimization time. This longer dedication to optimizing the calibration problem is one reason why the cascading surrogate method outperforms the RBFOpt optimization method in the controlled case study

with a much lower combined CVRMSE. In the real metered case study, the methods perform similarly, which we think stems from the fact the BEM is not configured accurately enough to let the calibration methods differentiate themselves from one another. In the following sections, we describe this problem in more detail and delve deeper into the individual trial results.

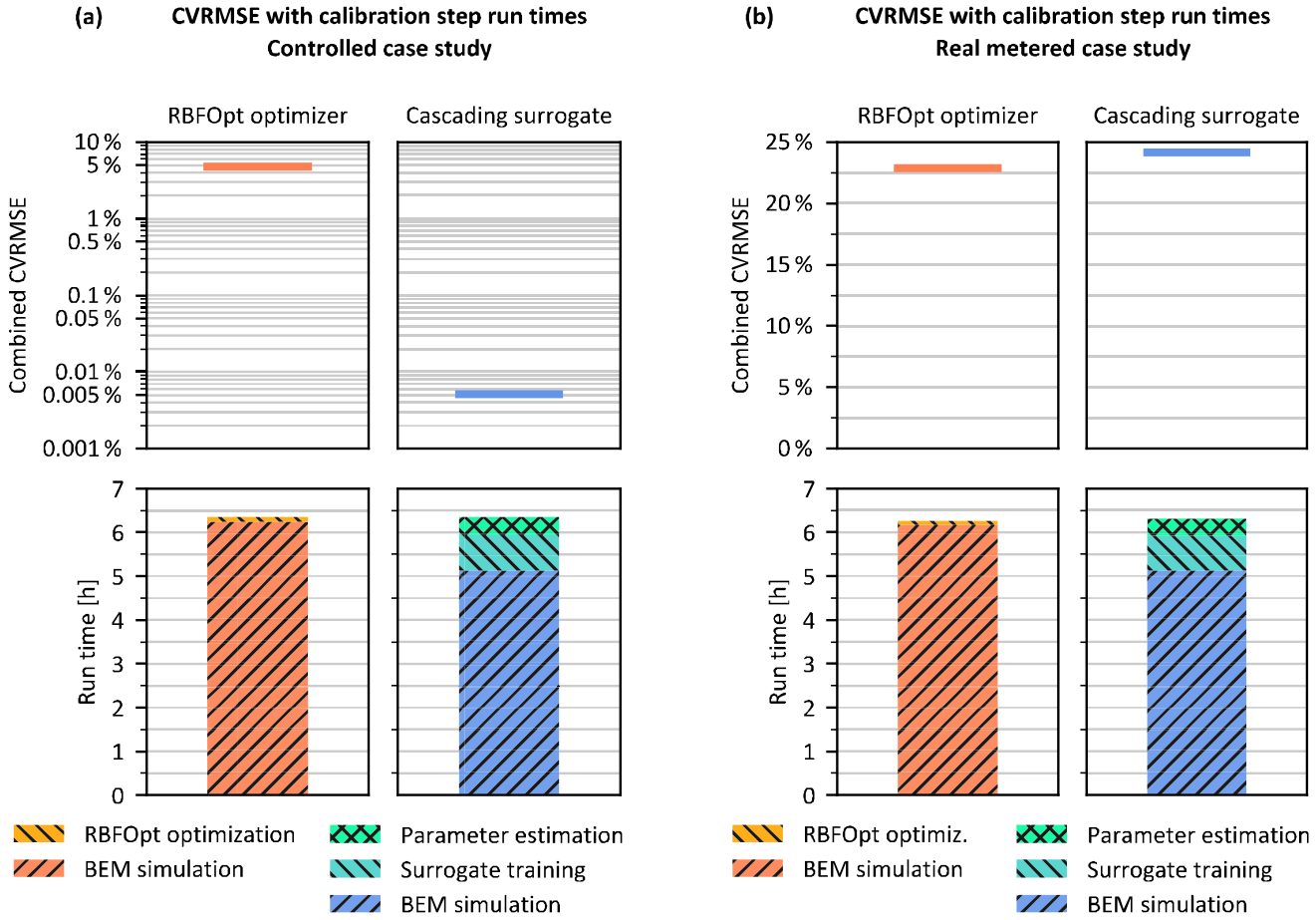

Figure 8: Comparing the combined CVRMSE and run times for the different calibration steps—averaged over 5 trials—for **(a)** the controlled case study and **(b)** the real metered case study.

3.5 Controlled case study

In Figure 9, we compare in more depth the calibration performance between the cascading surrogate method and the RBFOpt optimizer method. As previously mentioned, we ran 5 trials for each calibration method to account for some of the inherent randomness in BEM calibration.

We can see in Figure 9a that the final optimized parameter values when using the cascading surrogate calibration method (blue stars) always fall much closer to the actual parameter values (green lines) than the RBFOpt optimizer method (red dots). The final combined CVRMSE values for the cascading surrogate method are much lower than those of the RBFOpt optimizer method, with values between 0.0027 and 0.0075 % compared to between 3.6 and 6.4 % (Figure 9b). The parameter values of the cascading surrogate calibration method also have much lower variance than the RBFOpt method. The cascading surrogate calibration method therefore greatly outclassed the powerful RBFOpt optimizer in this controlled case study. The superior performance can be attributed to the efficacy of cascading design spaces and the accuracy of the surrogate model.

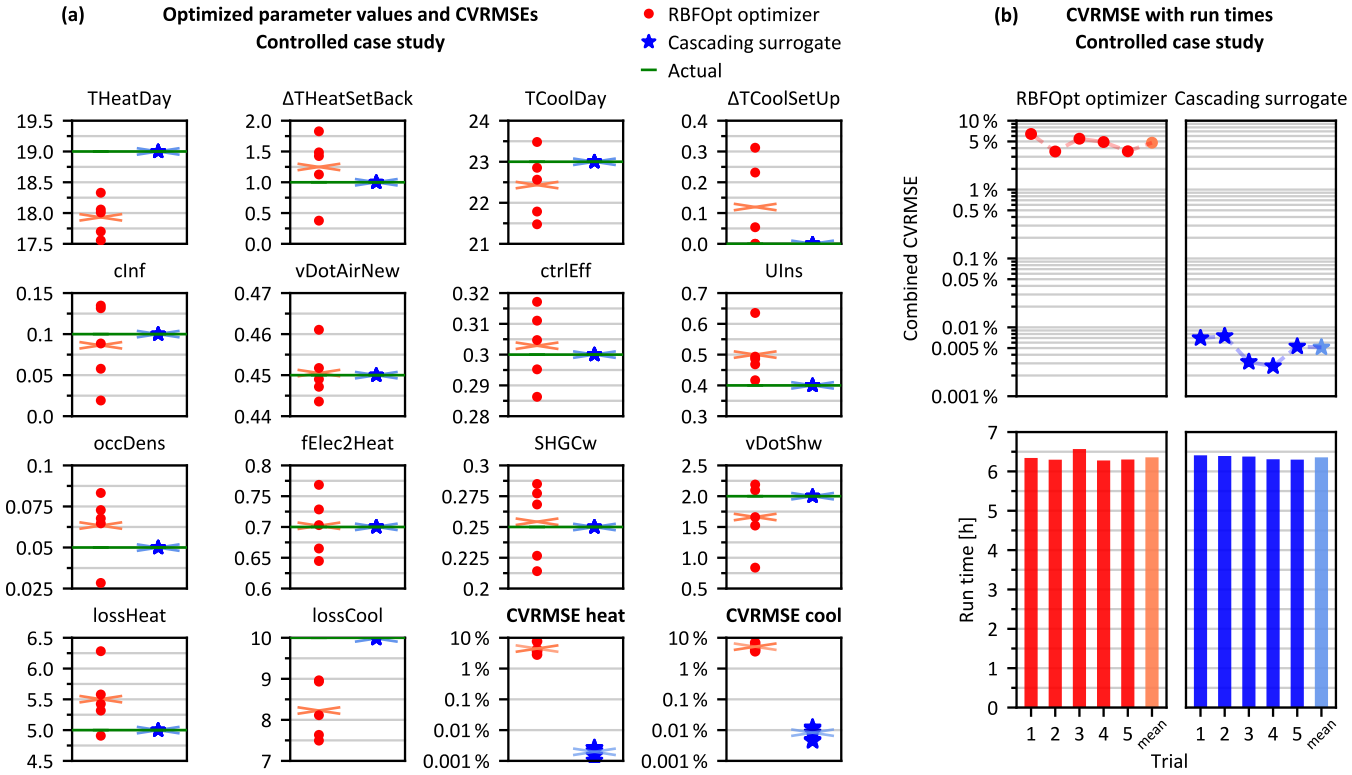

Figure 9: Comparing **(a)** the optimized parameter values and **(b)** the computation run times between the RBFOpt optimizer and the cascading surrogate method in the controlled case study.

To show how the bounds are increasingly restricted in the cascading surrogate calibration method, we present in Figure 10 the evolution of the bounds and the unknown building parameter estimates for the trial that is closest to the average final calibration result (trial #5—see Figure 9b). We can clearly see how the bounds become more and more restricted and how the estimates get closer and closer to the actual value of the unknown building parameters (*i.e.*, global minimum) thanks to the smaller distance over which the MPR surrogate model must interpolate when training. The mean squared error in the bottom right plot of Figure 10 is calculated with Equation (11) by comparing the estimated unknown building parameter values and the controlled values (see Table 4).

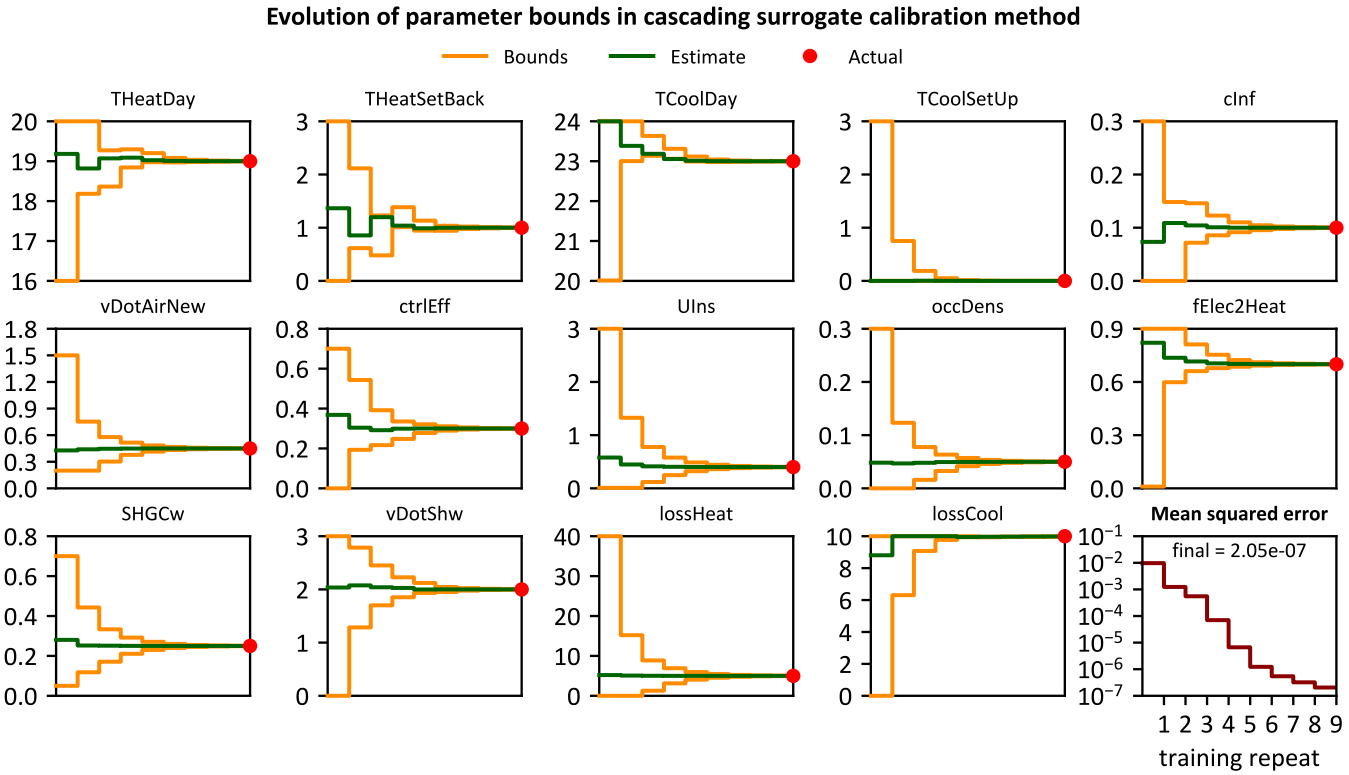

Figure 10: Showing how the surrogate model is trained in increasingly restricted unknown building parameter spaces, which zeros-in on the global minimum

3.6 Real Metered Case Study

In Figure 11, we compare the cascading surrogate calibration method and the RBFOpt optimizer in a real metered case study using the measured energy of White Hall from 2019. Throughout its 5 trials, the RBFOpt optimizer found combined heating and cooling CVRMSE values that are slightly lower than the cascading surrogate method, with values between 22.7 and 23.1 % compared to between 23.7 and 24.8 %. One possible reason for the poorer performance of the cascading surrogate method with real metered data is that the calibration method's hyperparameters were tuned with simulated testing data. These tuned hyperparameter values (especially how fast and how much the parameter bounds are reduced) might not be the most optimal for calibrating a BEM with more noisy, real metered data. Despite the poorer performance, the cascading surrogate method appears to have more reasonable calibrated values for some the unknown building parameters compared to the RBFOpt optimizer. For example, the calibrated heating temperature setpoint is more logically around 20 °C with the cascading surrogate method versus around 17 °C with the RBFOpt optimizer. We are thus led to question whether a lower CVRMSE necessarily indicates that the BEM is more calibrated. There are many local minimums in BEM calibration and some minimums appear more plausible to building energy modelers than others. In this case, the cascading surrogate calibration method's local minimums seem more plausible.

Yet, in both calibration methods, the values of other unknown building parameters are lower than expected, notably the fraction of electricity consumption that is turned into heat (fElec2Heat), the solar heat gain coefficient of the windows (SHGCw), and the maximum occupancy density (occDens). Although the low fElec2Heat value is not easily explainable, the low SHGCw value could be explained by a fairly high window-to-wall ratio and the closure of the blinds for much of the year. With respect to occupancy, the maximum occupancy density is nearly zero in both methods, which is not realistic since there are indeed people in the building. It is well known that occupant behavior is very difficult to measure accurately and is one of the chief reasons for the discrepancy between simulated and measured data in BEM calibration (Azar and Menassa 2012). The metered electricity consumption was used to inform the simplified occupancy schedule used in the simulations (refer to our previous article (Herbinger, Vandenhof, and Kummert 2023)); however, it is most certainly still a gross estimation of the actual occupancy schedule, thus the optimizers deemed that the occupancy in the building was actually increasing the BEM's calibration error and decided to zero-out the occupancy density parameter.

In the end, even though the combined hourly CVRMSE in both methods is below ASHRAE's guideline of 30% (Table 1), the BEM does not appear to be that well calibrated when considering the unusual building parameter values. It is likely that both calibration methods optimized the unknown building parameters as much as they could before getting stuck in local

minimums due to a poorly configured BEM. Even though we attempted to model the BEM as accurately as possible, these unusual building parameter values indicate that we simply did not know enough about the HVAC systems and the controls/occupancy schedules to correctly model the case study building of White Hall. In order to use the cascading surrogate calibration method to evaluate the efficacy of energy retrofit measures (ECMs), detailed building audits would need to be carried out to learn more about White Hall and more accurately configure the BEM. In a controlled case study with no BEM modelling error, the cascading surrogate calibration method performed exceptionally well. Therefore, we can expect that if a more accurately configured BEM is used in this real metered case study, our method would find a lower error between simulated and measured energy and more believable building parameter values.

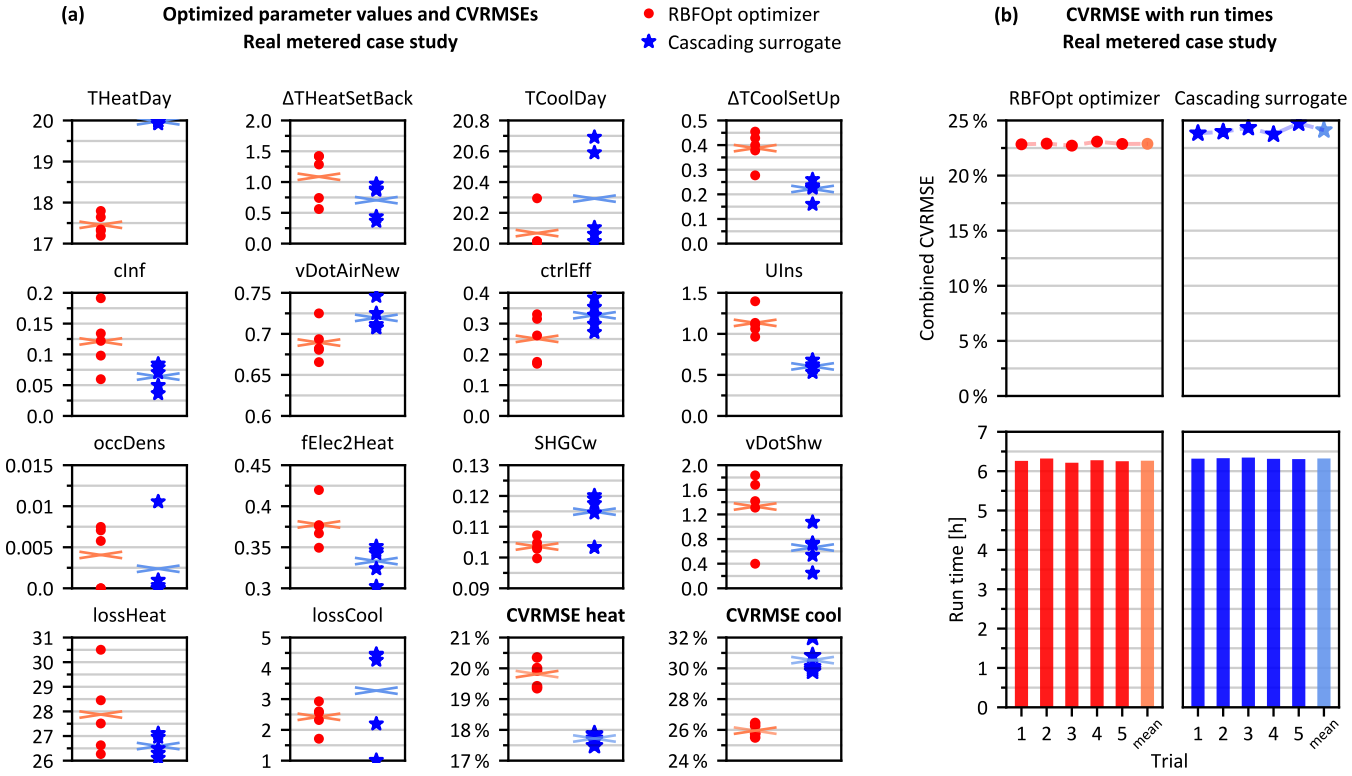

Figure 11: Comparing **(a)** the optimized parameter values (points), mean (cross), and standard deviations (lines) and **(b)** the computation run times between the RBFOpt optimizer and the cascading surrogate method in the real metered case study.

4 Conclusion

Typically, in the building energy model (BEM) calibration literature, surrogate models of BEMs are trained a single time over the entire unknown building parameter space. The surrogate model is then used to find the unknown building parameter values that minimize the difference between the surrogate model's prediction and the measured energy consumption of the modeled building. The surrogate is trained on a large range of unknown building parameter values and hence must interpolate between points that are far away to minimize prediction error. We developed a novel BEM calibration method where the surrogate model interpolates over smaller and smaller distances, thereby improving its ability to predict the measured energy consumption. Instead of training a surrogate model once over the entire unknown building parameter space, we retrain the surrogate model multiple times over increasingly restricted unknown building parameter spaces. This cascading surrogate calibration method follows four steps: (1) generate input (unknown building parameters) to output (simulated heating and cooling consumption) training data over the entire unknown building parameter space; (2) use this data to train a surrogate model in a supervised way; (3) estimate the unknown building parameter values that minimize the difference between the surrogate model's output and the measured energy consumption; and (4) restrict the bounds of the unknown building parameters around these estimates and repeats steps (2) to (4) a predefined number of times. The final unknown building parameter estimates after these repetitions are the calibrated values.

The design spaces are restricted around the previous estimated unknown building parameter values, which appear to become closer and closer to the global minimum thanks to the smaller distance over which the surrogate model must interpolate. However, this positive feedback loop towards the apparent global minimum is only possible if the surrogate model emulates

the BEM relatively accurately at the very first training repetition. For this, we developed an accurate multiple polynomial regression (MPR) surrogate model based on 8760 cubic ridge regressions, one for each hour of the year. The MPR surrogate model is both accurate and fast, since it fits a flexible, lightweight, non-linear function for each hour of the year separately, thereby accounting for the different exterior (*i.e.,* weather) and interior variables that impact the building's energy consumption at each hour of the year.

We compared the performance of this cascading surrogate calibration method to a popular calibration method in the literature: coupling a black box optimizer to the detailed BEM. We used the powerful RBFOpt black box optimizer, which has proven to have the best performance among open-source optimizers in Wortmann's (2019) benchmarking study of various building performance simulation tasks. We compared the calibration methods in a controlled and real metered case study of White Hall, an academic building on the Cornell University campus. In the controlled case study, the "true" energy consumption was the hourly heating and cooling consumption of a randomly selected configuration of the White Hall BEM. The interest of this case study is that the calibrated values of the unknown building parameters are known so the accuracy of the calibration methods can be assessed. In the real metered case study, the true energy consumption is the hourly measured heating and cooling consumption during 2019, allowing us to test the robustness of the calibration methods on real building data.

In the controlled case study, the cascading surrogate calibration method greatly outperformed the RBFOpt optimizer, finding combined hourly heating and cooling CVRMSEs of between 0.0027 and 0.0075 % throughout 5 trials, while the RBFOpt optimizer found CVRMSEs of between 3.6 and 6.4 % throughout its 5 trials. In the real metered case study, the RBFOpt optimizer this time found slightly lower CVRMSEs of between 22.7 and 23.1 % compared to between 23.7 and 24.8 %. for the cascading surrogate method. One possible reason for the poorer performance of the cascading surrogate method with real metered data is that the calibration method's hyperparameters were tuned with simulated testing data. These tuned hyperparameter values (especially how fast and how much the parameter bounds are reduced) might not be the most optimal for calibrating a BEM with more noisy, real metered data. However, given the fact that both methods performed quite similarly and that the calibrated values of the unknown building parameters were sometimes not very plausible, it is very likely that the White Hall BEM was not modeled accurately enough to allow the two calibration methods to really differentiate themselves.

No matter the robustness of the calibration method, no method can overcome certain modelling errors of the BEM. Either the correct degrees of freedom are not modelled (*e.g.*, incorrect HVAC system configuration) or there are not enough degrees of freedom (not enough parameters to tune). Yet, in a controlled case study where there is no modelling error, the novel cascading surrogate calibration method performed very well, outclassing a powerful black box optimizer, indicating that it can be very effective at BEM calibration.

5 Future Work

To provide further evidence of the cascading calibration method's performance, it should be validated with other buildings and on other measured data, such as whole-building electricity consumption. Furthermore, the hyperparameters of the cascading calibration method should be tuned with real measured data, as opposed to simulated data, to see if it improves the method's performance.

6 Data availability

Data will be made available on request.

7 Declaration of interest

We declare that we have no conflicts of interest.

8 Acknowledgments

We would like to thank the National Science and Engineering Research Council of Canada, the Fonds de recherche du Québec Nature et Technologie, the Arbour foundation, the Fondation et alumni de Polytechnique Montréal, and Hydro-Québec for providing funding in the form of scholarships to the first author.

9 References

ASHRAE. 1975. *Standard 90-1975: Energy Conservation in New Building Design*. Atlanta, GA, USA: American Society of Heating, Refrigerating and Air-conditioning Engineers.

ASHRAE. 2014. *Guideline 14-2014, Measurement of Energy and Demand Savings*. Atlanta, GA, USA: American Society of Heating, Refrigerating and Air-conditioning Engineers.

Azar, Elie, and Carol C. Menassa. 2012. "A Comprehensive Analysis of the Impact of Occupancy Parameters in Energy Simulation of Office Buildings." *Energy and Buildings*, Cool Roofs, Cool Pavements, Cool Cities, and Cool World, 55 (December):841–53. https://doi.org/10.1016/j.enbuild.2012.10.002.

Beal, Logan DR, Daniel C. Hill, R. Abraham Martin, and John D. Hedengren. 2018. "Gekko Optimization Suite." *Processes* 6 (8): 106. https://doi.org/10.3390/pr6080106.

Bellman, Richard E. 1957. *Dynamic Programming*. Princeton, N.J., USA: Princeton University Press.

Cant, Kevin, and Ralph Evins. 2022. "Improved Calibration of Building Models Using Approximate Bayesian Calibration and Neural Networks." *Journal of Building Performance Simulation* 0 (0): 1–17. https://doi.org/10.1080/19401493.2022.2137236.

Chen, Jianli, Xinghua Gao, Yuqing Hu, Zhaoyun Zeng, and Yanan Liu. 2019. "A Meta-Model-Based Optimization Approach for Fast and Reliable Calibration of Building Energy Models." *Energy* 188 (December):116046. https://doi.org/10.1016/j.energy.2019.116046.

Chong, Adrian, Yaonan Gu, and Hongyuan Jia. 2021. "Calibrating Building Energy Simulation Models: A Review of the Basics to Guide Future Work." *Energy and Buildings* 253 (December):111533. https://doi.org/10.1016/j.enbuild.2021.111533.

Coakley, Daniel, Paul Raftery, and Marcus Keane. 2014. "A Review of Methods to Match Building Energy Simulation Models to Measured Data." *Renewable and Sustainable Energy Reviews* 37 (September):123–41. https://doi.org/10.1016/j.rser.2014.05.007.

Cowan, John. 2002. "International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings - Vol 1." Washington, DC, USA: U.S. Department of Energy.

Ferrara, Maria, Ciro Lisciandrello, Alessio Messina, Mauro Berta, Yufeng Zhang, and Enrico Fabrizio. 2020. "Optimizing the Transition between Design and Operation of ZEBs: Lessons Learnt from the Solar Decathlon China 2018 SCUTxPoliTo Prototype." *Energy and Buildings* 213:109824. https://doi.org/10.1016/j.enbuild.2020.109824.

Gilks, Walter R., Sylvia Richardson, and David Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London, U.K.: Chapman and Hall/CRC.

Herbinger, Florent, Colin Vandenhof, and Michaël Kummert. 2023. "Building Energy Model Calibration Using a Surrogate Neural Network." *Energy and Buildings* 289 (June):113057. https://doi.org/10.1016/j.enbuild.2023.113057.

Li, Guangchen, Wei Tian, Hu Zhang, and Xing Fu. 2023. "A Novel Method of Creating Machine Learning-Based Time Series Meta-Models for Building Energy Analysis." *Energy and Buildings* 281 (February):112752. https://doi.org/10.1016/j.enbuild.2022.112752.

Li, Wancheng, Zhe Tian, Yakai Lu, and Fawei Fu. 2018. "Stepwise Calibration for Residential Building Thermal Performance Model Using Hourly Heat Consumption Data." *Energy and Buildings* 181 (December):10–25. https://doi.org/10.1016/j.enbuild.2018.10.001.

Monetti, Valentina, Elisabeth Davin, Enrico Fabrizio, Philippe André, and Marco Filippi. 2015. "Calibration of Building Energy Simulation Models Based on Optimization: A Case Study." *Energy Procedia*, 6th International Building Physics Conference, IBPC 2015, 78 (November):2971–76. https://doi.org/10.1016/j.egypro.2015.11.693.

Nagpal, Shreshth, Caitlin Mueller, Arfa Aijazi, and Christoph F. Reinhart. 2019. "A Methodology for Auto-Calibrating Urban Building Energy Models Using Surrogate Modeling Techniques." *Journal of Building Performance Simulation* 12 (1): 1–16. https://doi.org/10.1080/19401493.2018.1457722.

Østergård, Torben, Rasmus Lund Jensen, and Steffen Enersen Maagaard. 2018. "A Comparison of Six Metamodeling Techniques Applied to Building Performance Simulations." *Applied Energy* 211 (February):89–103. https://doi.org/10.1016/j.apenergy.2017.10.102.

Reddy, T.A., I. Maor, S. Jian, and C. Panjapornporn. 2006. "Procedures for Reconciling Computer-Calculated Results with Measured Energy Data." Technical Report. Atlanta, GA, USA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Tang, Boxin. 1993. "Orthogonal Array-Based Latin Hypercubes." *Journal of the American Statistical Association* 88 (424): 1392–97. https://doi.org/10.2307/2291282.

Tang, Boxin, and Devon C. Lin. 2015. “Latin Hypercubes and Space-Filling Designs.” In *Handbook of Design and Analysis of Experiments*, edited by Angela Dean, Morris Max, John Stufken, and Derek Bingham. New York, NY, USA: Chapman and Hall/CRC.

U.S. Department of Energy. 2015. “M&V Guidelines: Measurement and Verification for Performance-Based Contracts Version 4.0.”

Westermann, Paul, and Ralph Evins. 2019. “Surrogate Modelling for Sustainable Building Design – A Review.” *Energy and Buildings* 198 (September):170–86. https://doi.org/10.1016/j.enbuild.2019.05.057.

Wilde, Pieter de. 2014. “The Gap between Predicted and Measured Energy Performance of Buildings: A Framework for Investigation.” *Automation in Construction* 41 (May):40–49. https://doi.org/10.1016/j.autcon.2014.02.009.

Wortmann, Thomas. 2019. “Genetic Evolution vs. Function Approximation: Benchmarking Algorithms for Architectural Design Optimization.” *Journal of Computational Design and Engineering* 6 (3): 414–28. https://doi.org/10.1016/j.jcde.2018.09.001.

Yang, Tao, Yiqun Pan, Jiachen Mao, Yonglong Wang, and Zhizhong Huang. 2016. “An Automated Optimization Method for Calibrating Building Energy Simulation Models with Measured Data: Orientation and a Case Study.” *Applied Energy* 179 (October):1220–31. https://doi.org/10.1016/j.apenergy.2016.07.084.

Yuan, Jun, Victor Nian, Bin Su, and Qun Meng. 2017. “A Simultaneous Calibration and Parameter Ranking Method for Building Energy Models.” *Applied Energy* 206 (November):657–66. https://doi.org/10.1016/j.apenergy.2017.08.220.