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

Scalable Bayesian Optimization for Model Calibration: Case Study on Coupled Building and HVAC Dynamics

Chakrabarty, Ankush; Maddalena, Emilio; Qiao, Hongtao; Laughman, Christopher R.

TR2022-030 April 06, 2022

Abstract

Model calibration for building systems is an key step to achieving accurate and reliable predictions that reflect the dynamics of real systems under study. Calibration becomes particularly challenging when integrating building and HVAC dynamics, due to large-scale, nonlinear, and stiff underlying differential algebraic equations. In this paper, we describe a framework for calibrating multiple parameters of coupled building/HVAC models using scalable Bayesian optimization (BO), whose advantages include global optimization without requiring gradient information, and its ability to perform calibration in a data-efficient manner. The proposed methodology is improved online via two additional steps: domain tightening and domain slicing, both of which leverage the surrogate calibration cost function. We demonstrate effectiveness of the proposed algorithm by simultaneously calibrating 17 parameters (including emissivities, heat transfer coefficients, and thickness of walls/floors) of a Modelica model of joint building and HVAC dynamics, with 2 weeks worth of training data. This high-dimensional calibration task is solved via our proposed method, which yields parameters that are $> 90\%$ accurate with < 1000 model simulations, and the outputs of the final calibrated model on unseen testing data complies with standard ASHRAE calibration guidelines.

Energy and Buildings 2022

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

Scalable Bayesian Optimization for Model Calibration: Case Study on Coupled Building and HVAC Dynamics

Ankush Chakrabarty[a,1], Emilio Maddalena[b], Hongtao Qiao[a], Christopher Laughman[a]

[a] *Mitsubishi Electric Research Laboratories, Cambridge, MA, United States*
[b] *École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

Abstract

Model calibration for building systems is an key step to achieving accurate and reliable predictions that reflect the dynamics of real systems under study. Calibration becomes particularly challenging when integrating building and HVAC dynamics, due to large-scale, nonlinear, and stiff underlying differential algebraic equations. In this paper, we describe a framework for calibrating multiple parameters of coupled building/HVAC models using scalable Bayesian optimization (BO), whose advantages include global optimization without requiring gradient information, and its ability to perform calibration in a data-efficient manner. The proposed methodology is improved online via two additional steps: domain tightening and domain slicing, both of which leverage the surrogate calibration cost function. We demonstrate effectiveness of the proposed algorithm by simultaneously calibrating 17 parameters (including emissivities, heat transfer coefficients, and thickness of walls/floors) of a Modelica model of joint building and HVAC dynamics, with 2 weeks worth of training data. This high-dimensional calibration task is solved via our proposed method, which yields parameters that are $> 90\%$ accurate with < 1000 model simulations, and the outputs of the final calibrated model on unseen testing data complies with standard ASHRAE calibration guidelines.

Keywords: Parameter estimation, Gaussian processes, Machine learning, Bayesian methods, Sensitivity analysis, Digital twins

1. Introduction

Physics-informed simulation models and digital twins of heating, ventilation, and cooling (HVAC) systems play a critical role in predicting system dynamics and enabling analysis, control, and optimization of buildings and equipment [1, 2]. Key advantages of physics-based modeling are that designs can be based upon geometric and material information readily available from construction documents, and that the information encoded in their mathematical structure tends to demonstrate accurate extrapolative and predictive properties in comparison to more generic model structures. These predictive capabilities often come at the

*Email: chakrabarty@merl.com. Phone: (+1) 617-758-6175.

cost of increased nonlinearity and numerical stiffness, which can make these models difficult to simulate, and often impossible to obtain closed-form solutions or analytical (equation-based) representations.

Calibration mechanisms are essential to enable precise predictive performance of a simulation model for a given building, as the initial parameters obtained from tables of physical properties or architectural drawings may deviate from the actual materials or geometry used in the construction process, while other model parameters (e.g. heat transfer coefficients) may only be derived from correlations or empirical observations. The parameter values that most accurately represent observed data must therefore be systematically identified by algorithms that optimize a given calibration-cost map, such as the mean squared error. Once identified, the calibrated parameters of physics-based models can be useful not only for generating accurate simulations of the system dynamics, but also because the parameters themselves could have physical meaning and, therefore, be used to infer information about the system state.

Maps from model parameters to calibration-cost often present numerical challenges because they tend to be quite nonlinear, they are not always differentiable or convex, and their parameter sensitivities often vary over a wide range of magnitudes. Furthermore, measured data is corrupted by environmental effects or process noise, limiting the effectiveness of gradient-based methods. Population-based, gradient-free searches are scalable and effective, but incur high computational expenditure as they require extensive simulations [3, 4], which renders them unsuitable for calibrating stiff dynamical models that require long simulation times. Solely relying on dynamical estimators such as Kalman filters can also limit calibration performance due to multi-rate dynamics and low generalizability of linearized state-space models typically used to design these estimators [5, 6].

Coupled interactions between dynamic systems, as with an HVAC system and its associated building envelope, should also be accounted for during calibration. Although calibration of each of these system models is often performed independently, a number of variables (such as room air temperature) contain information about both the envelope and the HVAC system, so that the calibration of the joint system has the potential to yield more accurate parameters than the independent calibration of each subsystem. *Decoupling calibration of the building envelope from the HVAC is a missed opportunity.* The calibration of this integrated model presents additional challenges that must be addressed through algorithm design. This is due to the larger model size and the fact that the joint system often has dynamics over more widely separated time constants than either of the subsystem models[1].

Probabilistic learning algorithms can be used construct an approximation of the parameter-to-calibration-cost map to guide model calibration for systems with integrated dynamics of buildings and HVAC equipment, while accounting for noise and uncertainty. The learner explores subregions of the parameter space with high uncertainty (exploration) or high likelihood of obtaining a better solution (exploitation) and queries

[1]We show evidence of these numerical complications in Fig. 3 later in the paper.

the model only where collecting simulation data is likely to yield useful information. Intelligent, iterative sampling leads to lower simulation data requirements compared to widely used calibration mechanisms [7]. Bayesian methods, including Bayesian calibration [8–10], are particularly effective because they provide uncertainty quantification capabilities, as they both incorporate prior knowledge one might have of the building or HVAC system at hand and provide confidence envelopes around the nominal predictions specifying its degree of certainty without requiring a large amount of data [11].

The most widely used Bayesian algorithms for calibrating dynamical models and energy models fall under the umbrella of Markov chain Monte Carlo (MCMC) methods [12–14]. While these methods provide demonstrably excellent solutions for building energy calibration with few unknown parameters, they suffer from three major limitations: they typically require a large number of iterations for 'burn-in' and exhibit slow convergence in high-dimensional parameter spaces [13], they require preconditioning steps such as sensitivity analysis which is itself computationally expensive, and they require learning of multi-input multi-output maps via probabilistic learning methods to replace the emulator. Our proposed BO methodology is related to, but distinct from, the philosophy of Bayesian calibration. Although both Bayesian Calibration (BC) and Bayesian Optimization (BO) aim at adjusting parameters of simulation models, important differences between the two methodologies exist, described next. Whereas BC explicitly accounts for the presence of states x and outputs y, which are later used to compute metrics such as the CVRMSE cost, BO bypasses the estimation of x and relies solely on the availability of the measurements y and the dataset of parameters θ with their associated calibration cost. As numerous simulation runs are typically required in BC, it is customary to design meta-models to replace the high-fidelity simulators [15]. On the other hand, BO relies directly on the high-fidelity simulator and reduces the number of times simulations need to be performed, thereby reducing the need for a large number of initial simulations for meta-modeling. Coupled with the previous point, an advantage of BO over BC methods is that we do not model the states or dynamics, we instead rely on a model of the cost, which is a simpler learning problem. Even though both approaches usually rely on Gaussian process models, the posterior distribution is often analytically intractable in BC and users have to resort to different flavors of MCMC to gather samples from it [16]. MCMC is notoriously sample-heavy and requires a large number of samples just for burn-in [17]. In contrast, the posterior density can be found in closed-form in BO, dispensing with the need of performing MCMC, and we can analytically evaluate the posterior for inference.

In this work, we demonstrate a proposed GP-based Bayesian optimization (GP-BO) method for calibrating grey-box models of joint building-HVAC equipment systems. Advantages of this method include: BO usually requires few iterations to converge, does not require prior sensitivity analysis since it automatically determines relevance of parameters within the training phase, and a multi-input one-output map is learned (which reduces learning complexity) since we only seek to approximate the calibration-cost function rather than replacing the more complex underlying physics-informed model, thereby preserving fidelity.

Furthermore, BO methods are model agnostic: since we do not assume knowledge of the underlying model equations, BO can be used to calibrate both white-box and black-box models. Irrespective, we assume that *the model's closed-form representation is unknown, i.e. it is black-box from the perspective of the calibration algorithm.*

Gaussian processes (GPs) have been previously shown to be particularly effective for calibrating building energy maps; see [18–20]. In all these case studies, however, the dynamics of the underlying system are not considered, and the number of parameters is quite small. As pointed out in [21], the computational stability and efficiency of GP-based methods scale poorly when the number of iterations of the calibration algorithm is large. This is because standard implementation of Bayesian optimization using exact GP has cubic-time and quadratic-storage complexity on the number of data-points [22], which prevents its application to calibrating a large number of parameters. In order to overcome this important limitation, a variety of approximation techniques are available under the unifying name of sparse Gaussian processes (SGPs) [23, 24]. These methods rely on summarizing the information of the original dataset with a smaller collection of representative data points, also called *inducing points* or *pseudo-inputs*, whose number is determined by the user. In doing so, the domain expert can trade-off the approximation precision and the computational complexity of the resulting GP surrogate (more inducing points = more precision and more complexity, and vice versa). In this paper, we employ a state-of-the-art SGP technique for scalability during calibration, that is theoretically sound, easy to implement using open-source software tools, and yields fast and accurate results in practice.

The **contributions** of this paper are as follows: (1) We study the problem of developing scalable algorithms for calibrating physics-informed dynamical models of joint building and equipment dynamics. *This paper is a first attempt at learning parameters of a complex, large-scale dynamical model that accurately reflects the physical and engineering processes involved at both the equipment-level and the building-level.* (2) We employ data-driven Gaussian processes (GPs) for learning a parameter-to-calibration-cost function that contains the true cost with high probability. The GP also generates confidence bounds around predicted function values that quantify the prediction uncertainty at various regions in the parameter space. *A major advantage of this is that the surrogate model is constructed from parameters to calibration-cost which is typically a lower-dimensional learning problem than learning large-scale nonlinear dynamics, which is typically how GPs are currently used [11, 12].* (3) We utilize confidence bounds to explore the parameter space without requiring numerous simulations by collecting simulation data for parameters in regions with large uncertainty bounds and high likelihood of containing global optima, both of which are estimated by designing an appropriate acquisition function (as per standard Bayesian optimization). *By leveraging the learned statistics to automatically trade-off exploitation and exploration, we reduce the total number of model simulations required to complete the calibration task, and do not require a large number of initial simulations for 'burning-in' a distribution.*

The rest of the paper is organized as follows. In Section 2, we describe the workflow of physics-based model calibration with BO. The section also provides an overview of Gaussian processes (GPs) and Bayesian optimization (BO). Section 3 explains how to make the BO framework scalable to a large number of parameters using sparse GP approximations, and discusses modifications to SGP-BO such as domain tightening and domain slicing that improves convergence speed. The potential of our proposed approach is demonstrated using Modelica simulations in Section 4, where 17 parameters of a physics-based dynamical model is calibrated to high accuracy. We also show that integrating the HVAC to building envelope dynamics resulted in sizeable model complexity, making the case study a significant test bed for calibration. We present our conclusions and open problems in the Section 5.

2. Model Calibration with Bayesian Optimization

An overview of the proposed calibration framework is presented in Figure 1. The following subsections describe in more detail the individual components illustrated in the figure.

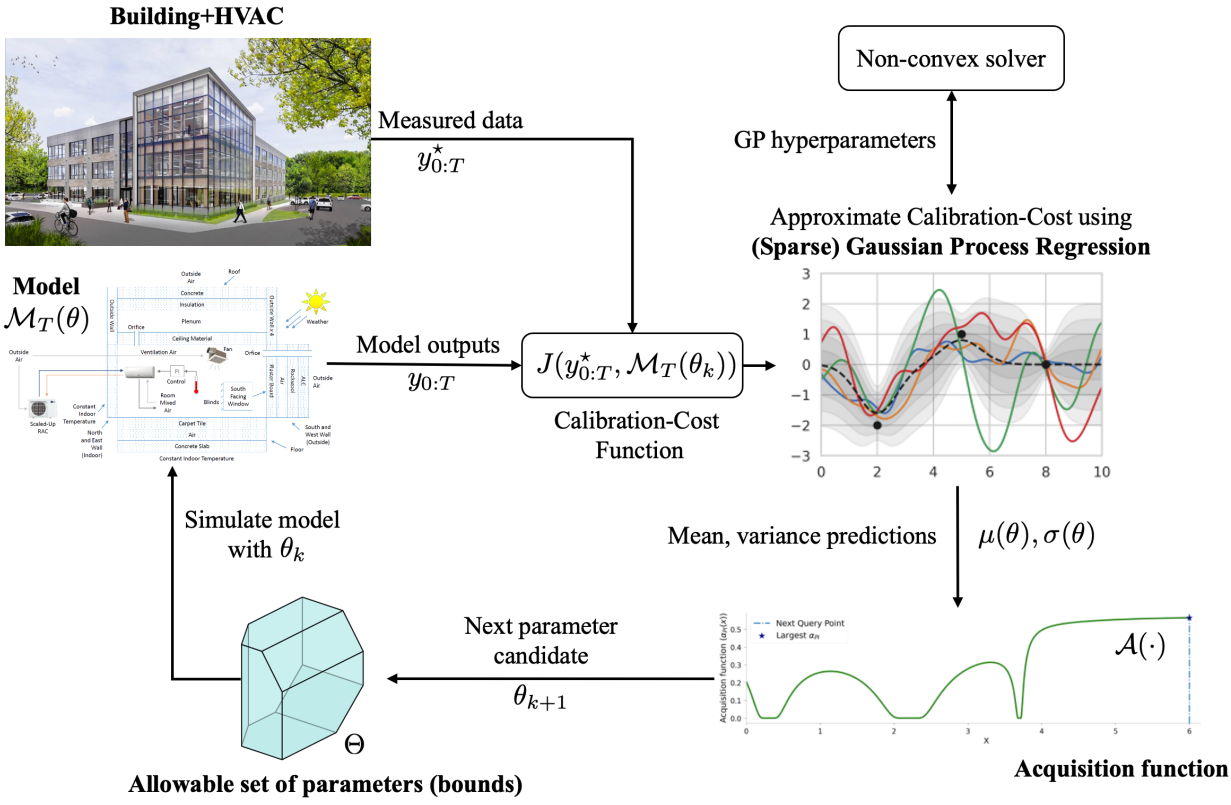

Figure 1: Schematic diagram of the proposed Bayesian optimization-based calibration method.

2.1. Model calibration

We denote by

$$y_{0:T} = \mathcal{M}_T(\theta)$$

an abstraction of a simulation model of the true system dynamics, parameterized by the parameter vector $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$. The admissible set of parameters Θ is assumed to be known at calibration time; for instance, Θ could denote a set of upper and lower bounds on parameters, obtained from physics or domain expertise. The output vector $y_{0:T} \in \mathbb{R}^{n_y \times T}$ contains all measured output data from the physical system over a time period $[0, T]$.

Note that we do **not** require an underlying mathematical description of the dynamics at design time. That is, $\mathcal{M}_T(\theta)$ could be a completely black-box or white-box model. We do require that the model can be evaluated for different parameters, and that simulating $\mathcal{M}_T(\theta)$ forward with a fixed (and admissible) set of parameters θ yields a time-sequence of outputs

$$y_{0:T} := \begin{bmatrix} y_0 & y_1 & \cdots & y_t & \cdots & y_T \end{bmatrix},$$

with each output measurement $y_t \in \mathbb{R}^{n_y}$.

This assumption implies that our method is applicable to a wide range of models proposed in the literature.

Example

For instance, consider the commonly used [14, 25] building energy model

$$y_t = \eta(x_t, \theta) + \delta(x_t) + \epsilon(x_t),$$

where η denotes the energy prediction, δ is the model discrepancy, and ϵ is the observation error. Clearly, by recursively simulating this model from $t = 0$ to $t = T$, one can obtain a representation that conforms to our abstracted model $\mathcal{M}_T(\theta)$.

Example

In the case of a state-space description [26] of the joint dynamics such as

$$\begin{aligned} \dot{x}_b &= f_b(x_b, x_e, u_b, w_b, \theta_b), \\ \dot{x}_e &= f_e(x_e, x_b, u_e, w_e, \theta_e), \\ y &= h(x_b, x_e, u_b, u_e, v_b, v_e), \end{aligned}$$

then defining $\theta := [\theta_b, \theta_e]$ and integrating from $[0, T]$ also yields a model of the form $\mathcal{M}_T(\theta)$. Here x, u, w, v denote states, inputs, process noise, and measurement noise, respectively, and the subscripts b and e correspond to the building and equipment, also respectively.

Since we are considering the problem of data-driven model calibration, we assume that we have some measured output data $y^\star_{0:T}$ that can be used to fit the model $\mathcal{M}_T(\theta)$. Our objective is thus to obtain the optimal set of parameters $\theta^\star$ such that the modeling error $y^\star_{0:T} - \mathcal{M}_T(\theta^\star)$ is minimized, according to a given distance metric. To this end, we propose an optimization problem to find the optimal parameters

$$\theta^\star = \arg\min_{\theta\in\Theta} J(y^\star_{0:T}, \mathcal{M}_T(\theta)). \tag{1}$$

While the designer is free to select any modeling error function J in (1), we have found

$$J(y^\star_{0:T}, \mathcal{M}_T(\theta)) := \log_{10}\left[\sum_{t=0}^{T}(y^\star_t - y_t)^\top W(y^\star_t - y_t)\right], \tag{2}$$

to work well, where W is a $n_y \times n_y$ positive-definite matrix that is used to assign importance or scale the output errors; recall $y_{0:T} = \mathcal{M}_T(\theta)$. The logarithm operator $\log_{10}(\cdot)$ promotes good numerical conditioning for learning by transforming very large or very small cost function values.

In order to perform global optimization, we solve the problem (1) by sampling the parameter space Θ, forward simulating the physics-based model $\mathcal{M}_T(\theta)$ over $[0, T]$ to obtain $y_{0:T}$, and computing the cost $J(y^\star_{0:T}, y_{0:T})$. By doing so, we avoid dependence on the underlying description of $\mathcal{M}_T(\theta)$, since just the simulated outputs are sufficient. Clearly, in high-dimensional parameter spaces, the number of samples required to obtain good solutions to (1) can be large unless the sampling is done intelligently. In this paper, we propose the use of Bayesian optimization to reduce sampling complexity by building probabilistic models of the mapping between the parameters and the calibration-cost, and iteratively exploiting the uncertainty associated with this probabilistic model. An overview of BO is presented next.

2.2. Overview of Bayesian optimization (BO)

The Bayesian optimization (BO) algorithm typically consists of two steps that automatically balance exploration and exploitation [27]. Probabilistic machine learning methods are used to approximate the map from parameters to the calibration-cost function J. Note that though the functional form of J is known from the model output error through (2), we do not know how J depends on θ, which is why the map $\theta \mapsto J$ needs to be learned. Representing the calibration cost as a stochastic process is a key difference from prior Bayesian calibration methods that focus on modeling the state transitions or building energy output directly. By quantifying the uncertainty from parameter to calibration-cost, we can indirectly model various uncertainty sources on the building model, along with noise, in a holistic manner without learning a high-dimensional dynamical map from multiple states to multiple outputs. By learning a stochastic representation, one can use the approximation to generate a predictive distribution for J at each parameter θ. Furthermore, this predictive distribution is used to generate subsequent search directions, with a focus on subregions of Θ where the function most likely contains the global solution $\theta^\star$ which minimizes the cost (1). After a new sample is acquired in the promising subregion, the probabilistic model is updated through Bayes

rule, thus incorporating new information and refining its predictions. The process is then repeated until a stopping criterion in met. Gaussian processes are the prevailing surrogate model choice in BO due to the existence of a closed-form model update expression as well as a closed-form objective to tune it [22]. Note that the final parameter candidate selected by BO is deterministic, even though the cost function approximation is stochastic.

We utilize GPs to model a stochastic process, equivalently, a distribution over functions. The underlying assumption made is that the calibration cost function J to be optimized has been generated from such a prior distribution, characterized by a zero mean and a kernelized covariance function $\mathcal{K}(\theta, \theta')$. The covariance function $\mathcal{K}$ is singularly responsible for defining the characteristics of the associated functions such as smoothness, robustness to additive noise, and so on. While many kernel functions are available, we have found (empirically) that the Matérn 3/2 function provides a good approximation of calibration-cost functions.

Assume that we have already evaluated the objective at N_θ input samples. Let this training data be denoted by

$$\{(\theta_k^D, J(\theta_k^D) + \nu_k\}_{k=1}^{N_\theta},$$

where $\nu_k \sim \mathcal{N}(0, \sigma_n^2)$ is additive white noise in the measurement channel with zero-mean and unknown covariance σ_n^2. After specifying a kernel function, one can compute the following elements

$$K_D(\theta) = \begin{bmatrix} \mathcal{K}(\theta, \theta_1^D) & \cdots & \mathcal{K}(\theta, \theta_N^D) \end{bmatrix}$$

and

$$\mathcal{K}_D = \begin{bmatrix} \mathcal{K}(\theta_1^D, \theta_1^D) & \cdots & \mathcal{K}(\theta_1^D, \theta_N^D) \\ \vdots & \ddots & \vdots \\ \mathcal{K}(\theta_N^D, \theta_1^D) & \cdots & \mathcal{K}(\theta_N^D, \theta_N^D) \end{bmatrix}.$$

With $K_D(\theta)$ and $\mathcal{K}_D$, we then define the GP predictive distribution.that is, the posterior, characterized by a mean function $\mu(\theta)$ and variance function $\sigma^2(\theta)$ given by

$$\mu(\theta) = K_D(\theta)^\top \mathcal{K}_n^{-1} J(\theta), \tag{3a}$$

$$\sigma^2(\theta) = \mathcal{K}(\theta, \theta) - K_D(\theta)^\top \mathcal{K}_n^{-1} K_D(\theta), \tag{3b}$$

with $\mathcal{K}_n = \mathcal{K}_D + \sigma_n^2 I$. The accuracy of the predicted mean and variance are strongly linked to the kernel selection and the best (in some sense) set of its *hyperparameters*. The latter are internal constants such as the length scale l, the vertical scale σ_0, and the noise variance σ_n^2. There are a variety of methods to optimize these hyperparameters, but the most common one consists in maximizing the log-marginal likelihood (MLE) function

$$\mathcal{L} = -\frac{1}{2} \log |\mathcal{K}_n| - \frac{1}{2} J(\theta)^\top \mathcal{K}_n^{-1} J(\theta) - \frac{p}{2} \log 2\pi. \tag{3c}$$

This is a widely adopted statistical objective whose maximum selects the model from which the observed data are more likely to have come. Although (3c) is non-convex, it can be solved using quasi-Newton methods or adaptive gradient methods [22]. If available, prior knowledge can be used to bias the estimation process towards values that the designer regards as being more sensible. This is can be easily incorporated in the same framework, and is referred to as maximum a posteriori (MAP) estimation. At this point, one has a GP model defined in (3a) and (3b), as well as a principled way of training it (3c).

The exploration-exploitation trade-off in BO methods is performed via an acquisition function $\mathcal{A}(\cdot)$. The acquisition function uses the predictive distribution given by the GP to compute the expected utility of performing an evaluation of the objective at each set-point θ. The next set-point at which the objective has to be evaluated is given by

$$\theta_{N_\theta+1} := \arg\max \mathcal{A}(\theta).$$

In this work, we use a lower confidence bound (LCB) acquisition function given by

$$\mathcal{A}_{\text{LCB}} = \mu(\theta) - \kappa\sigma(\theta), \tag{4}$$

where $\kappa \in \mathbb{N}$ is an integer usually ≥ 2. This acquisition function estimates the expected improvement of the steady-state power generated by the next set-point versus the current best solution. The maximum of this acquisition function may be computed efficiently by generating random samples on Θ, computing $\mathcal{A}_{\text{LCB}}$ for each sample, and choosing the sample maximum as the next set-point. As this function only depends on the GP approximated function and not on the actual objective J, the maximization of $\mathcal{A}(\cdot)$ involves computing (3) rather than expensive function evaluations. After a suitable number of iterations N_θ, the GP regressor is expected to learn the underlying function J and the best solution obtained thus far by the acquisition function is denote the best set of parameters for the model. The selection of N_θ is a design decision: it is usually informed by practical considerations such as the total number of simulations achievable within a practical time budget.

3. Scalable BO for High-Dimensional Parameter Spaces

3.1. Sparse Gaussian Processes for BO (SGP-BO)

Sparse Gaussian processes require $\{\theta'_k\}$, $k = 1, \ldots, M_{\theta'}$ as inducing points, also known as pseudo-inputs, to compress large datasets. Typically, these are chosen to be much fewer than the initial number of samples, that is, $M_{\theta'} \ll N_\theta$. Moreover, let $\mathcal{K}_{Dm}$ be the tall matrix of kernel evaluations at the inputs θ_k^D and inducing points $\{\theta'_k\}$, and $\mathcal{K}_{mD}$ be its transpose. Finally, $\mathcal{K}_{mm}$ denotes the matrix analogous to $\mathcal{K}_D$, but with evaluations at the inducing points—hence, $\mathcal{K}_{mm}$ is much smaller than $\mathcal{K}_D$.

The high computational complexity of classical GPs is due to computing the determinant and inverse of the $N_\theta \times N_\theta$ kernel matrix $\mathcal{K}_D$ in the likelihood loss (3c), which are needed both in the training and

prediction phases (see the expressions in (3)). To circumvent this problem, sparse techniques make use of a low-rank matrix

$$\tilde{\mathcal{K}}_D = \mathcal{K}_{Dm}\mathcal{K}_{mm}\mathcal{K}_{mD}, \quad \tilde{\mathcal{K}}_D \approx \mathcal{K}_D \tag{5}$$

along with the Woodbury inversion lemma and the Sylvester determinant theorem to greatly reduce the overall complexity; for more details, we refer to [23]. As an outcome of this process, the final SGP predictive mean and variance expressions only involve the inversion and the determinant of $M_{\theta'} \times M_{\theta'}$ matrices, which is much easier to compute. Besides the simplified matrix algebra, SGPs have a statistical interpretation: they are grounded in the independence of new and past data given the inducing points. For the calibration task, the surrogate calibration-cost learned based on inducing points form a bottleneck of approximation quality that require new candidate parameters to be sufficiently novel and informative to break through in order to be considered for evaluation [28].

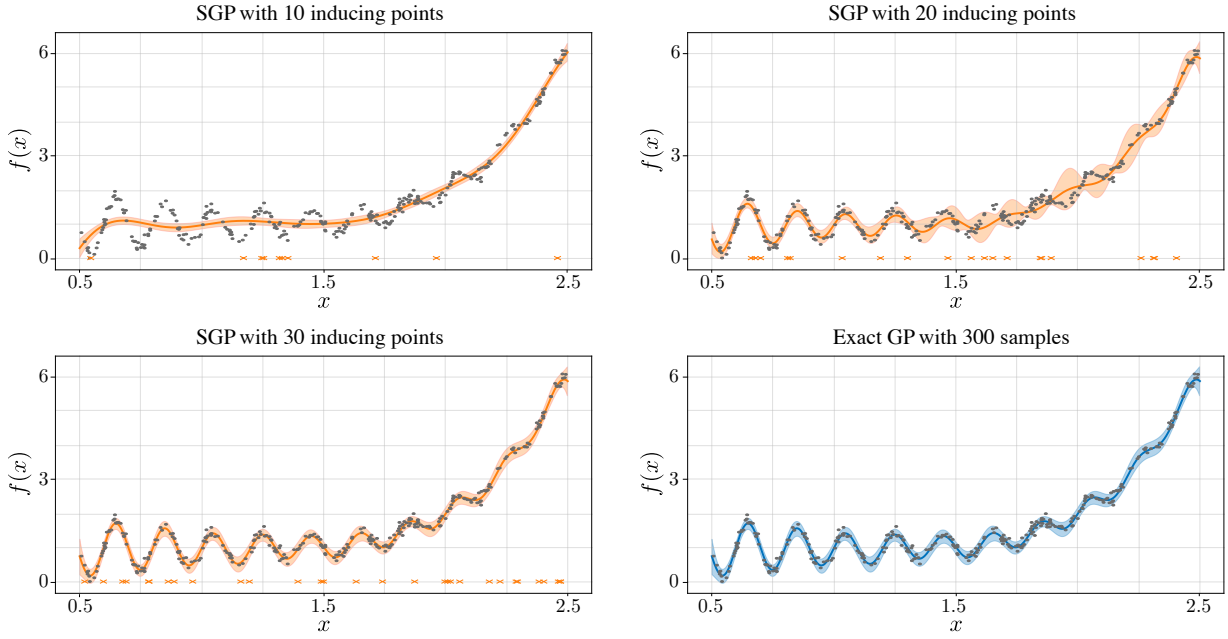

Figure 2: A comparison among sparse approximations employing an increasing number of inducing points, and the exact Gaussian process. The data-points are depicted as gray circles, whereas the inducing points are shown as orange crosses.

Among the various SGP methodologies found in the literature, the variational free energy (VFE) method stands out for its positive features: it is statistically consistent; it can only improve its performance when given more flexibility, i.e., more inducing points; and it is robust against overfitting as it fits the posterior density, never considering data directly [24]. The reader is referred to [23, 28] for two excellent reviews of SGPs and to [29] and [30] respectively for the original VFE work and a recent reinterpretation of it. The model itself is derived by minimizing the distance to the exact-GP posterior as measured by the Kullback-Leibler (KL) divergence metric, a common way of quantifying the mismatch between statistical distributions.

Using tools from variational calculus, a closed-form solution for this problem can be found, resulting in a (simplified) Gaussian process with the following predictive equations

$$\mu_{\mathrm{VFE}}(\theta) = \tilde{K}_D(\theta)^\top (\tilde{\mathcal{K}}_D + \sigma_n^2 I)^{-1} J(\theta),$$
$$\sigma^2_{\mathrm{VFE}}(\theta) = \mathcal{K}(\theta,\theta) - \tilde{K}_D(\theta)^\top (\tilde{\mathcal{K}}_D + \sigma_n^2 I)^{-1} \tilde{K}_D(\theta).$$

As explained in [24], training the VFE model is typically done through maximizing the following objective

$$\tilde{\mathcal{L}} = -\frac{1}{2}\log|\tilde{\mathcal{K}}_D + \sigma_n^2 I| - \frac{1}{2} J(\theta)^\top (\tilde{\mathcal{K}}_D + \sigma_n^2 I)^{-1} J(\theta) - \frac{1}{2\sigma_n^2}\mathsf{Tr}(\mathcal{K}_D - \tilde{\mathcal{K}}_D) - \frac{p}{2}\log 2\pi; \tag{6}$$

where $\mathsf{Tr}(\cdot)$ denotes the trace of a matrix. The maximization of $\tilde{\mathcal{L}}$ not only identifies the best kernel hyperparameters, but also optimizes the inducing point locations, hence completely specifying the sparse model. We chose VFE as our default sparse Gaussian process in this paper due to its superior performance in our numerical experiments along with a substantial computational speed-up.

The benefits of sparse GPs, specifically VFE, can be seen in Fig. 2 using a benchmark 1D function. The function is replicated using exact GP using 300 points. VFE approximations with an increasing number of inducing points are also presented for comparison. As can be seen from the plots, 10 inducing points did not suffice to learn a reasonable nominal model as most of the data fluctuations were neglected. With 20 inducing points, the SGP yielded better results, but certain regions of the domain between $x = 1.5$ and $x = 2.5$ presented low-quality predictions and a large uncertainty envelope. Finally, excellent results were obtained with 30 inducing points, being essentially indistinguishable from the exact GP with only one tenth of the number of effective inputs.

Pseudocode for the resulting SGP-BO algorithm is provided in Algorithm 1.

3.2. Incorporating Sensitivity Information

Most calibration mechanisms in the literature start with a sensitivity analysis on the parameter space to identify which parameters directly affect the model outputs, and how strongly. This sensitivity analysis is performed offline and typically requires access to an analytical mathematical model so that one can compute gradients and calculate sensitivity matrices [31]. In the absence of such an analytical model, one can generate sensitivity-like indices via sampling, but this requires a large number of offline samples, i.e., queries to the mathematical model [32]. Since BO methods explicitly construct GP models of the cost with uncertainty quantification, we can use these surrogates to learn the sensitivity (also called *relevance*) of each parameter, online, for free [33]. In this subsection, we propose two modifications to the SGP-BO algorithm that take parameter relevances into account to restrict the search domain (*domain tightening*) and slice the high-dimensional parameter space into tractable clusters of parameters, to promote scalability (*domain slicing*).

Algorithm 1 SGP-BO Algorithm

Require: Model, $\mathcal{M}_T(\theta)$

Require: Initial dataset $\mathcal{D}_1 = \{\theta, J\}_{0:N_{\mathsf{DT}}}$

Require: Acquisition Function, $\mathcal{A}$

Require: Inducing points, θ'

for $k = 1, 2, \cdots$ **do**

Perform min-max scaling of θ

Update the sparse Gaussian process model with $\mathcal{D}_k$

$\theta_{k+1} \leftarrow$ Find next best candidate, $\arg\max \mathcal{A}(\theta)$

$J(\theta_{k+1}) \leftarrow$ Compute cost by simulating model

$\mathcal{D}_{k+1} \leftarrow$ Augment dataset $\mathcal{D}_k \cup \{\theta_{k+1}, J(\theta_{k+1})\}$

end for

3.2.1. Domain Tightening

Algorithm 2 Domain Tightening

Require: Initial parameter space, Θ

Require: Top-ℓ_{DT} number, ℓ_{DT}

Require: Dataset so far, $\{\theta, J\}_{0:N_{\mathsf{DT}}}$

$\theta_{\ell_{\mathsf{DT}}} \leftarrow$ Find best ℓ_{DT} solutions by sorting J

$\Theta_{\mathsf{DT}} \leftarrow$ Compute range of each $\theta_{\ell_{\mathsf{DT}}}$

return Tightened parameter space, Θ_{DT}

Continue SGP-BO with tightened domain $\Theta \leftarrow \Theta_{\mathsf{DT}}$

The intuition behind domain tightening stems from the fact that after an initial exploration phase, the SGP-BO algorithm starts to exploit locations where good parameter sets likely reside, by sampling more densely in those subregions of the parameter space. After sufficient BO iterations, say a user-defined N_{DT}, we expect that the data $\{\theta_k^D, J_k^D\}_{k=0}^{N_{\mathsf{DT}}}$ obtained thus far contain a few parameter sets that generate low calibration costs. We can therefore use sorting to collect the top-ℓ_{DT} samples from the dataset that result in the ℓ_{DT} lowest calibration cost values, since calibration-costs are scalar-valued. Parameters associated with these top-ℓ_{DT} (top = smallest) cost values contain information about a subregion of the parameter space within which the optimizer most likely resides, and we can thus take the range of these top-ℓ_{DT} samples as the new parameter space. Since the simple regret of SGP-BO asymptotically decays to zero with increasing data [34], we expect that the range of the top-ℓ_{DT} samples will be subsets of the initial parameter space. This is why we refer to this method as 'domain tightening', which can be performed periodically to focus the search region of the parameters based on prior solutions. Note that if the calibration cost is locally

convex in a region around the optimal parameter set, then the range of the top-ℓ_{DT} parameters provides an indirect measure of sensitivity: indeed, the tighter the range is, the more sensitive the parameter. If not, then solutions in a wider range would admit the same low calibration cost values. A pseudocode for domain tightening is presented in Algorithm 2.

3.2.2. Domain Slicing

After a pre-defined number of BO iterations (equivalently, SGP-BO iterations) have been completed, one has a dataset with which to construct a surrogate GP (or SGP) model from the parameters to the calibration cost. We propose a method called domain slicing which leverages this surrogate model to perform sensitivity analysis/relevance determination on the parameter space of interest[2]. Specifically, one can slice the original n_θ-dimensional calibration problem into smaller chunks $n_{\theta_1}, n_{\theta_2}, \ldots$, such that they sum to n_θ by ranking and clustering the parameters by relevance. We expect that performing domain slicing will provide additional scalability by generating subproblems that are computationally tractable since they only calibrate a subset of the parameters that are most sensitive to the cost.

Algorithm 3 Domain Slicing

Require: Initial search domain, Θ

Require: Number of clusters, N_{cl}

Require: Dataset so far, $\mathcal{D} = \{\theta, J\}$

$\chi \leftarrow$ Perform relevance determination of all parameters

Cluster relevances into N_{cl} clusters

Select clusters whose centroids are highest relevance

Select parameters in high-relevance clusters

Generate new dataset with high-relevance parameters

Fix low-relevance parameters using optimizer in $\mathcal{D}$

Perform SGP-BO on high-relevance parameters

We begin by performing sensitivity analysis by following the algorithm described in [35], which ranks the parameters based on the variability of the SGP mean function in the direction of each parameter. Additionally, recall that $\{\theta'\}$ is the set of inducing points for the SGP, and let $\hat{J}$ denote the calibration cost estimates at these inducing points. Let $\mu = \mu(\theta')$ and $\Sigma = \sigma^2(\theta')$ denote the mean and variance functions, respectively, of the SGP model. At the k-th inducing point, the conditional distribution of the j-th parameter is given by

$$p(\theta_j | \theta'_k | - j) \sim \mathcal{N}(\tilde{\mu}_j, \tilde{\sigma}_j^2),$$

[2]If domain tightening has been performed, this parameter space is Θ_{DT}, otherwise Θ.

where

$$\tilde{\mu}_j = \mu_j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}(\theta_{-j} - \mu_{-j}),$$
$$\tilde{\sigma}_j^2 = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{j,-j}.$$

Here, the subscript notation j indicates the j-th row/column of the matrix is selected, and $-j$ indicates that the j-th row/column is excluded. For instance, $\Sigma_{j,-j}$ indicates the vector of Σ where all elements of the j-th row are taken except the j-th column element. Additionally, $\Sigma_{-j,-j}$ is the submatrix containing all elements of Σ except the j-th row and column. In order to derive a sensitivity measure, we compute the variance of the posterior mean along the j-th dimension, which is obtained by computing first and second moments of the conditional distribution $p(\theta_j|\theta'_k| - j)$, given by

$$\mathsf{Var}[\hat{J}_j^k] = \mathcal{I}_1 - \mathcal{I}_2^2,$$

where

$$\mathcal{I}_1 = \int_{\Theta_{\mathsf{DT}}} (\hat{J}_j^k)^2(\theta_j)\,\mathcal{N}(\theta_j|\tilde{\mu}_j, \tilde{\sigma}_j^2)\,\mathrm{d}\theta_j,$$
$$\mathcal{I}_2 = \int_{\Theta_{\mathsf{DT}}} (\hat{J}_j^k)(\theta_j)\,\mathcal{N}(\theta_j|\tilde{\mu}_j, \tilde{\sigma}_j^2)\,\mathrm{d}\theta_j.$$

As in classical sensitivity analysis [33], these integrations are typically done via gridding methods, and it is well-known that Gaussian integration is best approximated with Hermite polynomials/Gauss-Hermite quadrature. In particular, if we select N_h to be the order of the Hermite polynomials used for function approximation, we can rewrite $\mathcal{I}_1$ and $\mathcal{I}_2$ as

$$\mathcal{I}_1 \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N_h} \omega_i (\hat{J}_j^k)^2 (\sqrt{2}\tilde{\sigma}_j\varphi_i + \tilde{\mu}_j),$$
$$\mathcal{I}_2 \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N_h} \omega_i (\hat{J}_j^k)(\sqrt{2}\tilde{\sigma}_j\varphi_i + \tilde{\mu}_j),$$

where φ_i are the roots of the physicists' version of the Hermite polynomial, and ω_i are the corresponding weights. Once these summations have been evaluated for all dimensions and all $M_{\theta'}$ inducing points, we can compute the relevance using the average

$$\chi_j = \frac{1}{M_{\theta'}} \sum_{k=1}^{M_{\theta'}} \mathsf{Var}[\hat{J}_j^k].$$

The computational complexity of this method is $\mathcal{O}(n_\theta(M_{\theta'} + n_\theta^2))$ [35].

Upon computing the relevances/sensitivities $\{\chi\}_1^{n_\theta}$ of all parameters, we perform clustering (e.g., via K-means) and rank the clusters in the order of decreasing relevance. Since the least sensitive parameters do not strongly affect the calibration cost, we can fix these using the optimal solution found so far. We can

then perform SGP-BO for the parameters in the cluster or clusters that have high relevance. Since the most sensitive parameters will be a subset of the total parameter set, the corresponding calibration problem will require searching a lower-dimensional, or sliced, search domain. Pseudocode for domain slicing is provided in Algorithm 3.

4. Results and Discussions

4.1. Building and Equipment Model Description

The temporal behavior of the vapor-compression cycle is dominated by the heat exchangers over the time scales of seconds to hours, so the system models in this work used dynamic models of the heat exchangers and static (algebraic) models of the compressor and expansion valve. We assumed 1-D flow for the refrigerant so that properties only vary along the length of the pipes; we also assume that the refrigerant can be described as a Newtonian fluid, negligible viscous dissipation and axial heat conduction in the direction of flow, and negligible contributions to the energy equation from the kinetic and potential energy of the refrigerant. For the sake of simplicity, a lumped parameter method was used to characterize the dynamics of refrigerant flow in the heat exchangers. A multicomponent moist-air model was used for the air-side of this work, in which both dry air and water vapor were described by ideal gas equations.

A simple isenthalpic model was used for the electronic expansion valve, in which the mass flow rate is regularized in the neighborhood of zero flow to prevent the derivative of the mass flow rate from tending toward infinity. The flow coefficient is generally determined via calibration against experimental data [36].

The cycle models in this work included a variable-speed low-side scroll compressor, in which the motor is cooled by the low-pressure refrigerant entering the compressor. Due to the complex nature of the heat transfer and fluid flow through the compressor, we also used simplified 1-D models of this component to parsimoniously describe the system. The behavior of the compressor was described by relating the volumetric efficiency and isentropic efficiency to the suction pressure, discharge pressure, and compressor frequency. The compressor power consumption was also related to the compressor speed and the ratio of inlet and outlet pressures. The coefficients used for the functional forms of the volumetric efficiency and isentropic efficiency were also derived from experimental data [36].

Standard fan laws were used to describe the behavior of the heat exchanger fans. According to such models, the volumetric flow rate was assumed to be directly proportional to the fan speed, while the power consumed by the fan was assumed to be proportional to the cube of the fan speed. These simple algebraic models were scaled by experimentally measured values of fan speed, flow rate, and power; to minimize the error in these fits, linear and quadratic terms were also included in the power model to account for observed variations in the data.

The building models were based upon the open-source Modelica Buildings library [37], an extensive and well-tested library of components for the construction of dynamic building and building system models. The room model from the Buildings library is based on the physics-based behavior of the fundamental materials and components commonly used in the building construction industry. The zone air model incorporated into the room model is a mixed air single-node model with one bulk air temperature that interacts with all of the radiative surfaces and thermal loads in the room. The individual materials are parameterized by fundamental properties like thickness, thermal conductivity, and density, and can be combined and assembled into multi-layer constructions.

The building model consists of a one-story residence with nominal 2009 IECC-based construction, based on the model used in [38]. This residence has a floor area of 112.24 m^2 and is 2.6 m tall, and is oriented along the cardinal directions with a peak occupancy of 3 people per floor. Each exterior wall also has a window of 1.52 m by 2.72 m that admits solar heat gains into the spaces. A 10 cm thick concrete slab and 2 meters of soil below the house was also included to characterize interactions with the thermal boundary condition under the house, which was set to a constant 21°C. A peaked attic was also included with a maximum height of 1.5 meters, so that the building model includes two thermal zones.

Table 1 lists the parameters of the building and the heat pump selected to evaluate the efficacy of this new calibration method. Approximately equal numbers of parameters were selected from the envelope model and the cycle model to study the accuracy for each subsystem. These particular parameters were chosen because they are often difficult to measure in practice or to estimate from other physical quantities. For example, the refrigerant-side heat transfer coefficients (HTCs) depend on the amount of oil circulating in the pipe, the detailed configuration of tubes in the heat exchanger, and many other system and site-specific quantities. On the envelope side, the radiative emissivities in the IR and solar spectra were similarly selected because of their potential experimental variability. We then bounded the ranges of the parametric variation based on our field experience, though other ranges could be easily used.

To elucidate upon the complexity arising from integrating building and equipment, we linearize the Modelica model about an equilibrium state and analyse the resulting system state transition matrix. In Fig. 3, we show the sparsity pattern of the linearized system state transition matrix containing the states of the HVAC system dynamics (red), the building envelope dynamics (green), and the coupling between them (blue). The white space indicates zero values; it is clear that the state matrix is sparse. The 2-norm of the coupled state elements is 310.677, indicating that the elements are not trivially small: this indicates that removing the equipment model from the building and using a static/simplified proxy removes critical dynamics from the overall building system. From the eigenvalues, we can infer that the system is extremely stiff: the building eigenvalues (green stars) range within $[10^{-5}, 1]$ sec^{-1}, which implies that there are dynamics that act in the order of seconds to 10^5 seconds, which is approximately 1.15 days, which are reasonable time-spans for thermal dynamics in the building envelope e.g. wall temperatures. Conversely,

the HVAC dynamics are much faster and can range, using similar analysis, from the order of micro-seconds to the order of minutes. This is also expected since the pressure dynamics are extremely fast, but tube wall temperatures of the heat exchanger walls exhibit slower dynamics. Since the dynamics vary from milli-seconds to days, ODE solvers require very small step sizes for numerical integration, and simulations are performed over long time horizons to enable the observation of slow dynamics. The building dynamics alone are relatively better conditioned.

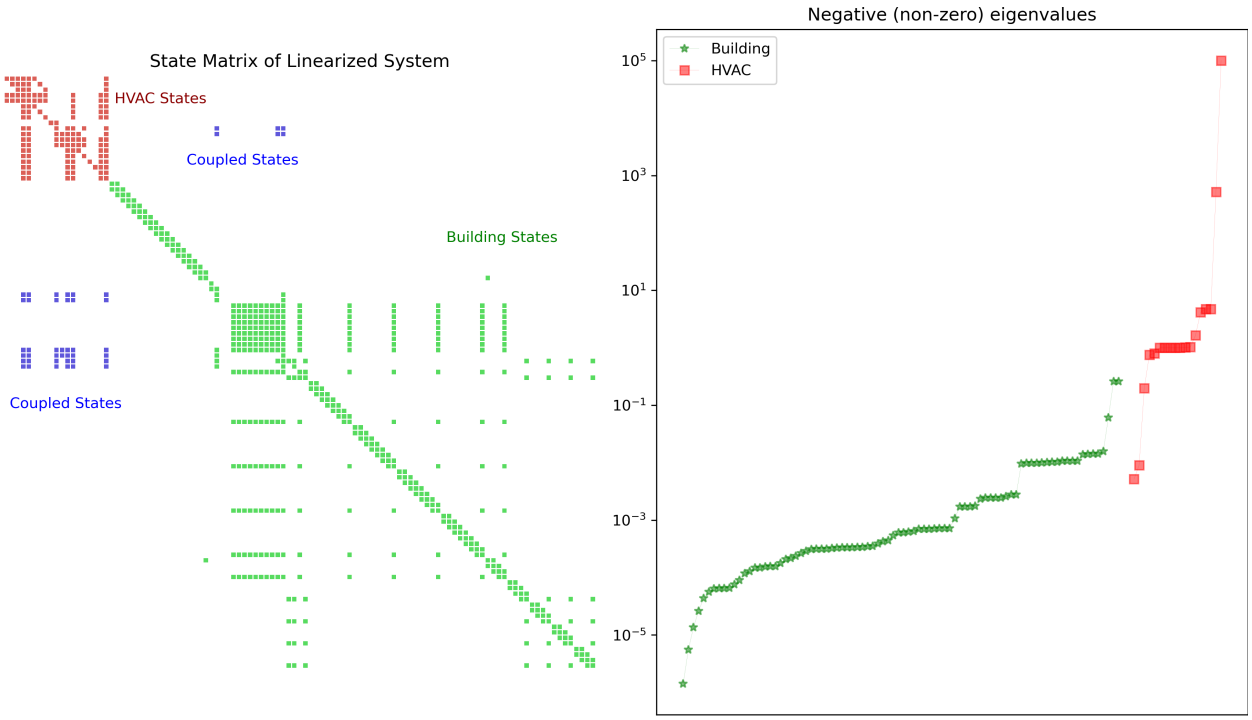

Figure 3: Sparsity pattern and eigenvalue spectrum of linearized HVAC and building dynamics.

4.2. Implementation Details

Modelica

This envelope model was connected to the cycle model, and a controller was implemented on the heat pump which used the compressor frequency to regulate the room temperature and the expansion valve position to regulate the evaporator superheat temperature. The controller also implemented anti-windup to maintain stability while enforcing minimum and maximum actuator limits. The resulting joint building envelope/HVAC model was simulated using the Atlanta-Hartsfield TMY3 file, and included convective and radiative heat loads of 2 W/m^2 and a latent load of 0.6 W/m^2 between the hours of 8 AM and 6 PM, with weather-driven disturbances outside of these hours. This model was exported from Modelica using the Functional Mockup Interface [39], and the resulting functional mockup unit (FMU) was imported into

Parameter	Variable	True Value	UB [%]	SGP-BO-DT	ACC [%]	SGP-BO-DTS	ACC [%]
Building Parameters							
Airflow infiltration rate	`VFlowExt`	3.368×10^{-2}	± 15	3.269×10^{-2}	97.1	-	-
Thickness of the floor	`xFloor`	1.016×10^{-1}	± 10	1.002×10^{-1}	98.6	-	-
Infrared emissivity of roof (out)	`IR-Roof-a`	9.000×10^{-1}	± 15	8.625×10^{-1}	95.8	8.863×10^{-1}	98.5
Solar emissivity of roof (out)	`Sol-Roof-a`	9.000×10^{-1}	± 15	9.352×10^{-1}	96.1	9.111×10^{-1}	98.8
Infrared emissivity of roof (in)	`IR-Roof-b`	7.000×10^{-1}	± 15	6.864×10^{-1}	98.1	-	-
Solar emissivity of roof (in)	`Sol-Roof-b`	7.000×10^{-1}	± 15	6.714×10^{-1}	95.9	-	-
Interior room air HTC	`hInt`	3.000	± 5	3.134	95.6	-	-
Exterior air HTC	`hExt`	1.000×10^{1}	± 5	9.678	96.8	-	-
HVAC Parameters							
Outdoor HEX HTC AF	`pfHTC-a`	1.000	± 15	9.374×10^{-1}	93.7	9.342×10^{-1}	94.3
Indoor HEX HTC AF	`pfHTC-b`	1.000	± 15	1.110	88.9	1.033	96.7
Indoor HEX Lewis number	`Le-a`	8.540×10^{-1}	± 5	8.625×10^{-1}	99.0	-	-
Outdoor HEX vapor HTC	`HTC-vap-a`	5.000×10^{2}	± 10	5.186×10^{2}	96.3	-	-
Outdoor HEX 2-phase HTC	`HTC-2ph-a`	3.000×10^{3}	± 10	3.251×10^{3}	91.6	-	-
Outdoor HEX liquid HTC	`HTC-liq-a`	7.000×10^{2}	± 10	7.380×10^{2}	94.6	-	-
Indoor HEX vapor HTC	`HTC-vap-b`	5.000×10^{2}	± 10	4.562×10^{2}	91.2	-	-
Indoor HEX 2-phase HTC	`HTC-2ph-b`	2.000×10^{3}	± 10	1.958×10^{3}	97.9	1.822×10^{3}	91.1
Indoor HEX liquid	`HTC-liq-b`	7.000×10^{2}	± 10	7.118×10^{2}	98.3	-	-

Table 1: Description of parameters, true values, and uncertainty expressed as a percentage from the true value after 1500 iterations of SGP-BO-DT and 500 subsequent iterations of SGP-BO-DTS. (HTC = heat transfer coefficient, HEX = heat exchanger, AF = adjustment factor, UB = uncertainty bound, ACC = accuracy)

Python using the FMPy package[3] to enable the application of the variety of machine learning tools available in that language. The inputs and outputs of this model were chosen to be similar to those which may be observed in an experimental setting. The inputs of the heat pump include the room temperature setpoint, the evaporator superheat setpoint, and the indoor and outdoor fan speeds. The inputs for the building envelope model include the convective, radiative, and latent heat loads as well as the weather variables provided in the TMY3 standard. These heat loads may be estimated to reasonable accuracy via occupancy detection, load surveys, or other similar methods.

In order to provide a meaningful testing environment for the proposed SGP-BO method, we constructed a testing model in Modelica that has the same envelope and HVAC system as the training model with the following key changes to produce non-trivially different dynamics. We changed the radiative and convective heat loads from 2 W/m^2 to 3 W/m^2 and the latent head loads from from 0.6 W/m^2 to 0.9 W/m^2. We also altered the occupancy start and end times of the building from 8AM–6PM to 6:30AM–3:30PM. The weather profile is changed from the Hartsfield Jackson, GA, USA airport to Charlotte Douglas, NC, USA airport via a change in the TMY3 weather file to generate different ambient temperatures and humidities. We also tested the proposed method for a longer simulation time, from 2 weeks of data for training to 4 weeks for testing, and in the month of August (testing) rather than July (training).

[3] `https://github.com/CATIA-Systems/FMPy`

GPyTorch

We implemented our sparse GP model in `GPyTorch`[4] in Python 3.8 (requires `PyTorch`) with a CPU [40], and chose the `Inducing` and `Matern` kernels with automatic relevance determination (ARD) so that each kernel hyperparameter is dimension-dependent. For training, the dataset was min-max scaled, so that each parameter is scaled to $[0, 1]$: this greatly improved matrix conditioning. Furthermore, we used an Adam optimizer for 2000 epochs per training of SGP hyperparameters, with a learning rate between 0.01 and 0.05.

Model Stiffness

We used a timeout in the simulator to stop the simulation after 300 s of CPU time, and excluded that parameter combination from the training set of the learner and the admissible parameter space. We chose 300 s as our time-out because we noticed that only certain poor choices of parameters resulted in increased stiffness of the underlying dynamical equations and required over 2000 s to simulate, eventually ending in oscillatory outputs. We implemented a time-out at 300 sec to reduce the amount of wasted simulation time, since good choices of parameters resulted in simulations that terminated successfully within 200 s. Of course, if we simulated for longer, or for other models, one would have to carefully select this time-out based on empirical evidence regarding the duration of good simulations.

4.3. Parameter Estimation Performance

We simulated the Modelica model for 14 days of July with the parameters of the model set to their true values (see Table 1) to collect ground-truth data for calibration. The 8 measured output sequences $y^{\star}_{0:T}$ of the model were collected at 5 minute intervals, and, unlike other methods [14], we do not require splitting the data into categories like weekdays or weekends.

Hereafter, we use GP and SGP interchangeably, since we used SGP as our learner. We initialized the SGP by choosing 100 randomly selected parameter samples from within the bounds Θ associated with each parameter (see Table 1). With each of these initial parameter samples, we simulated the Modelica model for the same time interval as the ground truth and obtained the estimated output sequence $y_{0:T}$. Subsequently, we evaluated the cost function (2) for each of the initial samples with $y_{0:T}$ and true outputs $y^{\star}_{0:T}$. This initial collection of parameters and calibration-cost values was used to construct the initial training set of the SGP. We used Matérn 3/2 kernels with dimension-wise separate length-scales, since the admissible parameter space is not normalized. We constructed the GP using the Python library `GPyTorch`, and used 500 epochs of an Adam solver to obtain the optimal hyperparameters for training. Unlike MCMC methods that require tens of thousands of iterations to converge, we set our BO method to run for 750 iterations, that is, the Modelica model is simulated 750 (BO iterations) + 100 (initial) = 850 times from $[0, T]$. We

[4] `https://gpytorch.ai/`

selected the acquisition function to be a lower-confidence-bound (4) with $\kappa = 1.96$. For acquisition function maximization, we adopted a uniform random sampling approach with 10,000 samples. This sampling is cheap since it only requires evaluation of the SGP, rather than the simulation model. The specific SGP framework used is the VFE method with 100 inducing points.

Table 1 shows that the best estimates of the parameters after 750 iterations are quite close to the true parameter values. Indeed, 13 of the 17 parameters (6/8 building and 7/9 equipment) are captured at $>$ 90% accuracy[5], despite using only one 2-week dataset and no additional pre-processing such as sensitivity analysis or data splitting. It is noteworthy that the 4 parameters with the lowest fits, highlighted in gray in this table, were all $> 85\%$. Given the lack of heat transfer from the surfaces affected by these parameters, the relatively poor quality of these fits matches our intuition for the low sensitivity of the measured outputs to these parameters.

Recall that an advantage of our proposed approach is that a good guess for initialization or a burn-in period to obtain a good prior distribution on the parameters is not required to acquire good parameter candidates. Instead the BO framework requires knowledge of the parameter ranges, which implies that the prior distribution over parameters is uniform (initially, no parameter set is assumed to be more likely than any other). Despite the lack of prior knowledge, due to the sequential nature of BO, the calibration performance is correlated with the initial SGP model. We therefore tested our proposed approach for robustness to initial conditions by running the calibration mechanism 50 times with different initial random seeds, so that different samples were extracted for the initial GP construction. The results of these simulation studies are shown in Fig. 4. The median (horizontal orange line), quartiles (horizontal box lines), and range (whisker ends) for the best parameter set obtained over 50 runs are shown using boxplots, with the true parameter value shown with a '$\star$'. We deduce from these plots that the best parameter estimates are close to their true values, with (predictably) the worst calibration performance exhibited by the inner roof parameters. Interestingly, the liquid heat transfer coefficients do not exhibit significant decline over runs, but the Lewis number does. It is likely that this variation can be attributed to the time varying moisture removal rate of the evaporating heat exchanger and the dependence of the indoor relative humidity on both ambient conditions and internal latent loads.

4.4. Effect of Domain Tightening and Comparison with An Existing Method

In Fig. 4, we show that our proposed SGP-BO-DT algorithm with domain tightening outperforms (over 20 runs, with DT every $N_{\mathsf{DT}} = 50$ iterations, $\ell_{\mathsf{DT}} = 50$ top candidates) the SGP-BO algorithm despite the same number of BO iterations and all GP hyperparameters remaining equal at the initial iteration. While the SGP- BO shows an initial decay of the calibration cost, the mean and 95% confidence intervals in this

[5] Accuracy of the k-th parameter θ_k is computed by $100 \times (1 - |\theta_k^{\text{true}} - \theta_k^{\star}|/\theta_k^{\text{true}})$.

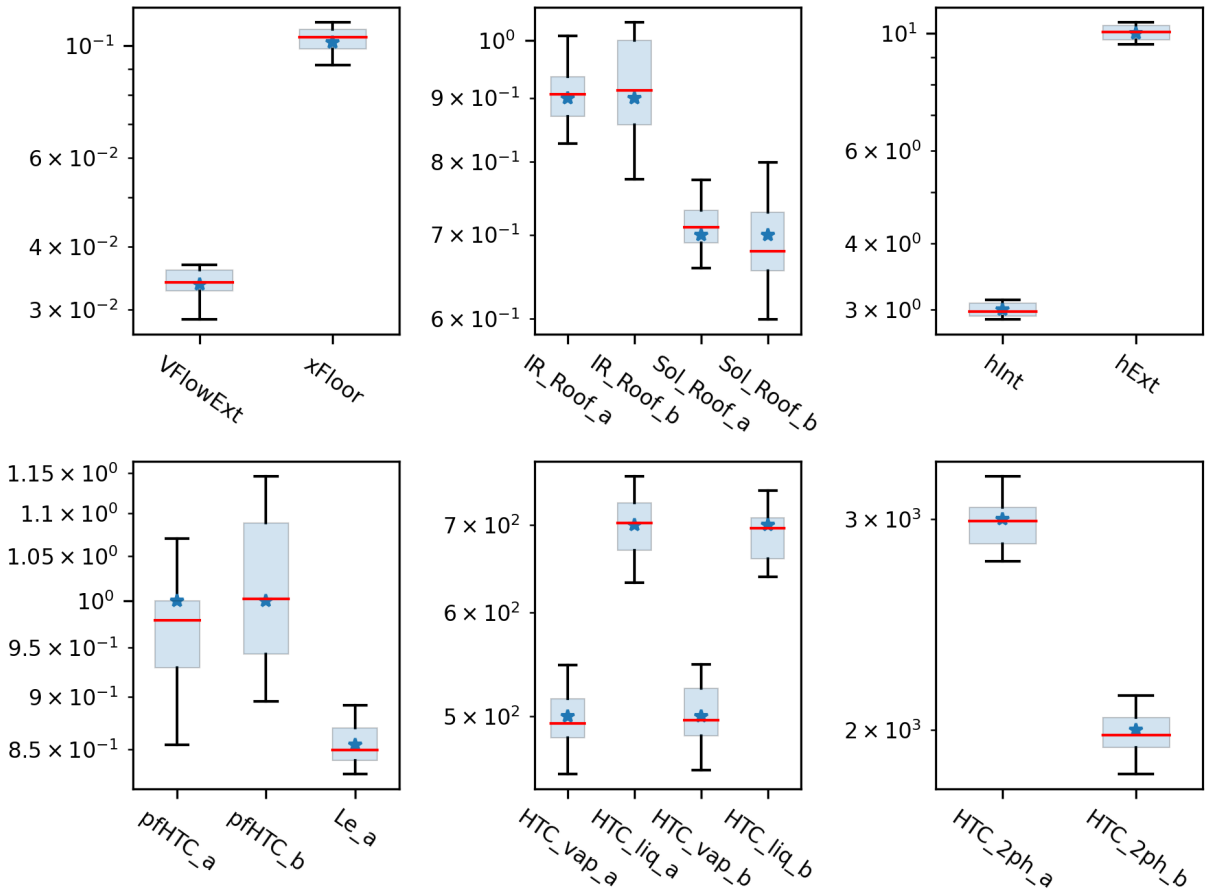

Figure 4: Illustration of robustness of SGP-BO-DT to random initialization. Boxplots of median (red line) and interquartile range (boxes) of all SGP-BO-DT computed parameters along with true values (star).

figure illustrate the fact that the domain tightening effect quickly reduces the relevant search space and promotes the generation of candidates that are more likely to reside in regions containing an extremal point. This results in both a faster decay of calibration cost and a better final set of candidates.

We also make a comparison against a version of the existing Bayesian calibration (BC) methodology, similar to the approach of [41]. BC approaches the calibration problem from a different set of modeling assumptions, and therefore, it does not make a fair comparison to compare the two approaches. In our case study, we have hundreds of states, which is beyond the state-size considered by BC algorithms which leverage MCMC sampling in much lower-dimensional spaces. We also provide references from which it is evident that our BO approach is much more sample efficient. For example, in [16], two case studies were carried out with 2 and 5 calibration parameters, requiring respectively a total of 791 and 503 training samples; our case study comprises 17 parameters and required 750 samples. BC and BO are not direct competitors, but calibrate quantities starting from a different set of assumptions and thus have also different sampling

requirements. To this end, we make some modifications to BC to ensure a fair comparison; we try to follow the key steps of the BC algorithm. Instead of iteratively re-learning the surrogate model as in BO, we adopt the BC method of constructing one GP model after acquiring an initial set of samples. After this GP is learned, we sample via MCMC methods from its posterior distribution to obtain a set of likely optimizers. We ensure that the total number of initial samples and MCMC samples is 850, the same as that of BO, to make the comparison fair. As expected, from Fig. 5, we observe that allowing the SGP to update as in BO, rather than keeping it constant as in BC, results in BO outperforming BC over most of the 20 runs. SGP-BO-DT comprehensively outperforms BC.

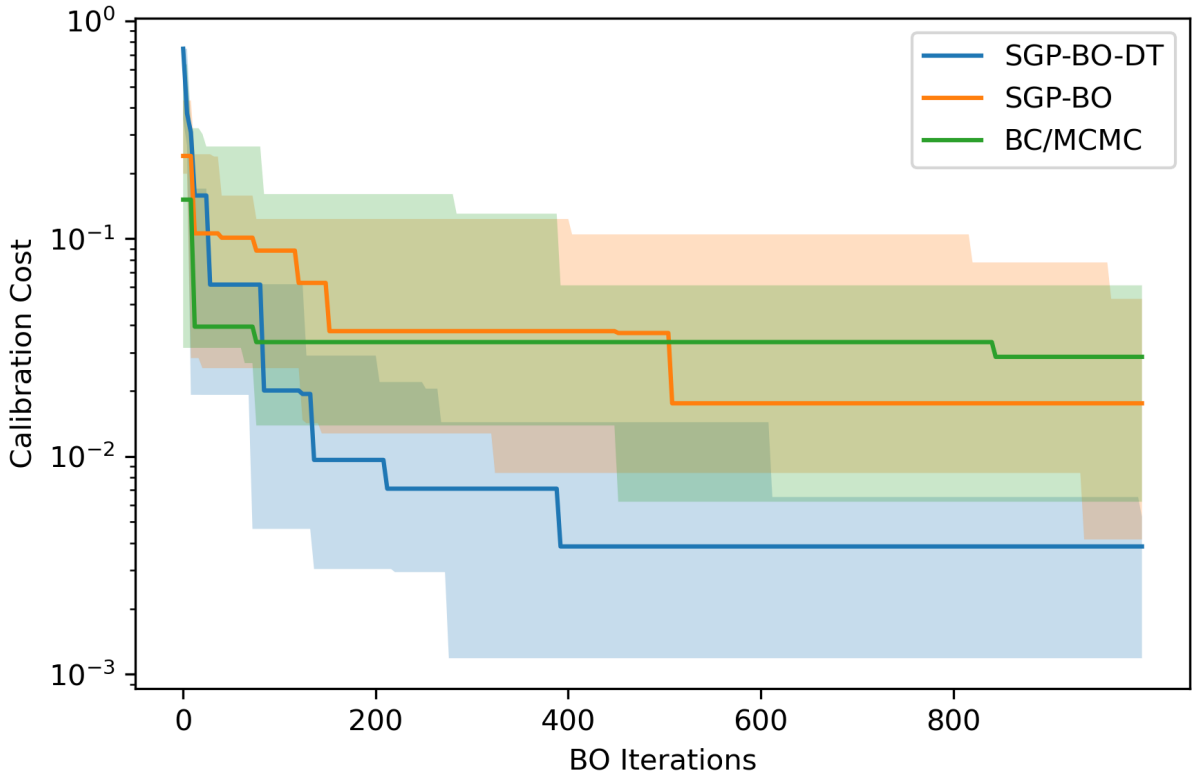

Figure 5: Comparison of SGP-BO with and without domain tightening against traditional MCMC-based Bayesian calibration.

4.5. Sensitivity Analysis and Domain Slicing

At the end of 1000 SGP-BO-DT iterations, we perform sensitivity analysis of the parameter set, as described in Section 3.2.2. The purpose of adding this step is to illustrate that the calibration procedure can be further improved by leaving out insensitive parameters, thereby reducing the effective search dimension. From the χ values shown in Fig. 6, we determine that the top-4 sensitive clusters include the solar and infrared emissivities of the outer roof, the adjustment factors for the heat transfer coefficients of the indoor

and outdoor heat exchangers, and the heat transfer coefficient of the two-phase indoor heat exchanger. The results can be justified from a physical point of view. IR-Roof-a affects the infrared radiation exchanged between the exterior surface of roof and the environment, whereas Sol-Roof-a affects the amount of solar radiation absorbed by the roof. For a given surface emissivity, a roof is more effective in losing or gaining heat radiation than a wall because the roof sees more of the sky than the wall. In this sense, these two parameters are the most crucial ones that determine the heat gains of the indoor air from the sun and the environment through the building envelope. On the other hand, the other three parameters (pfHTC-b, pfHTC-a and HTC-2ph-b) are among the most critical ones that determine the cooling performance of the HVAC equipment. In other words, the combination of these five parameters largely determine the net heat gain or loss of the indoor air and hence the room temperature.

Keeping only these 5 parameters as decision variables, and fixing the others according to the best candidates obtained so far (see Table 1 SGP-BO-DT column), we perform SGP-BO-DTS to obtain much tighter estimates of the emissivities. Of the five coefficients that are re-calibrated, we note that pfHTC-b improves in accuracy from 89% to over 96%, whereas the accuracy of the least sensitive parameter amongst the five, HTC-2ph-b, decreases from 98% to 91%. This is to be expected since the SGP-BO-DTS algorithm does not necessarily guarantee improvement parameter-wise, but only in the cost function decay. The new set of parameters obtained using SGP-BO-DTS results in a lower calibration cost compared to SGP-BO-DT by searching over a lower-dimensional subspace of Θ.

4.6. Predictive Performance

As per ASHRAE Guideline 14, a CVRMSE of $< 15\%$ indicates a good model fit with acceptable predictive capabilities [42]. In order to illustrate our calibration performance, we report the CVRMSE and the NMBE[6] metrics [43] in Table 2 for each of the model outputs. All of the parameters respect the ASHRAE guidelines in terms of the CVRMSE metric, showing the potential of our calibration mechanism and modeling approach. The highest CVRMSE is exhibited by the suction superheat, which can likely be attributed to sharp peaks produced in the signal related to large changes in the compressor speed and valve position during the rapid increase in the load caused by the morning solar load and the presence of occupants. Both training and testing scenarios are shown with noisy true data and model predictions in Fig. 7 and 8, respectively.

An additional benefit of this physics-informed approach to model calibration can be seen in Figure 9, as the calibrated model characterizes the building power consumption from July 16 to September 14 quite accurately, despite the fact that the building power signal is not used explicitly for calibration. The model error metrics for this scenario are provided in Table 2, which indicate that the error of the power signal is

[6] CVRMSE (Coefficient of Variation of the Root Mean Square Error) measures the variability of the errors between measured and simulated values. NMBE (Normalized Mean Bias Error) is the normalized average of the error sequence. We refer the reader to [43] for their mathematical definitions.

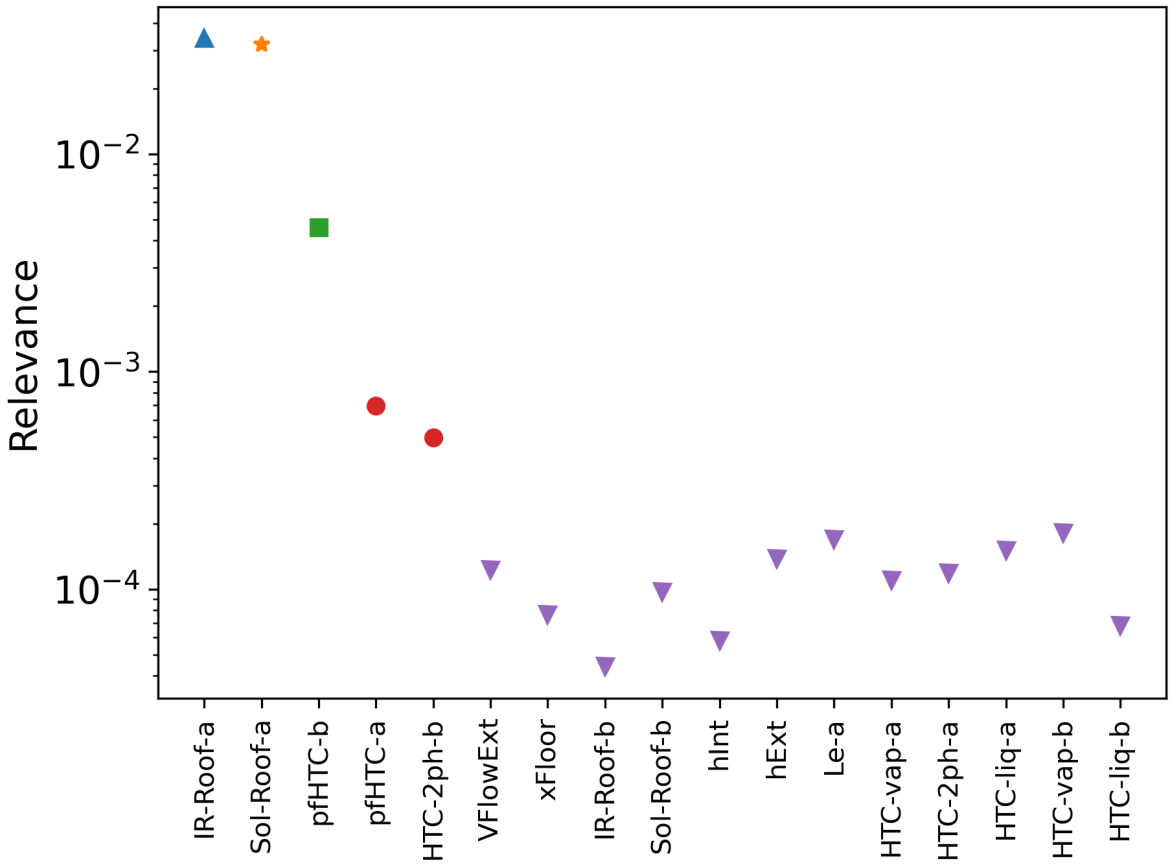

Figure 6: Relevance of parameters. The different symbols indicate clusters of varying relevance. The vertical axis is logarithmic.

within ASHRAE guidelines, and this level of accuracy is further corroborated by the low errors between the true and estimated power over most days in this study. While the absolute error is below 10% for most of this simulation, there are few days in which this error exceeds 20% due to the fact that the ambient temperature is near the limits of the range seen during training: either near 290 K or 310 K. Due to the scarcity of training data at these temperatures, as well as a jump in the room temperature due to actuators in the HVAC equipment reaching their limits, the quality of the model predictions at these points in time is lower than the average. Nevertheless, the calibrated model generally exhibits excellent predictive performance in conditions not seen during training.

5. Conclusions and Future Work

In this paper, we developed a Bayesian optimization methodology for calibrating physics-based models with fewer simulations than would be required for existing methods by learning the calibration-cost map

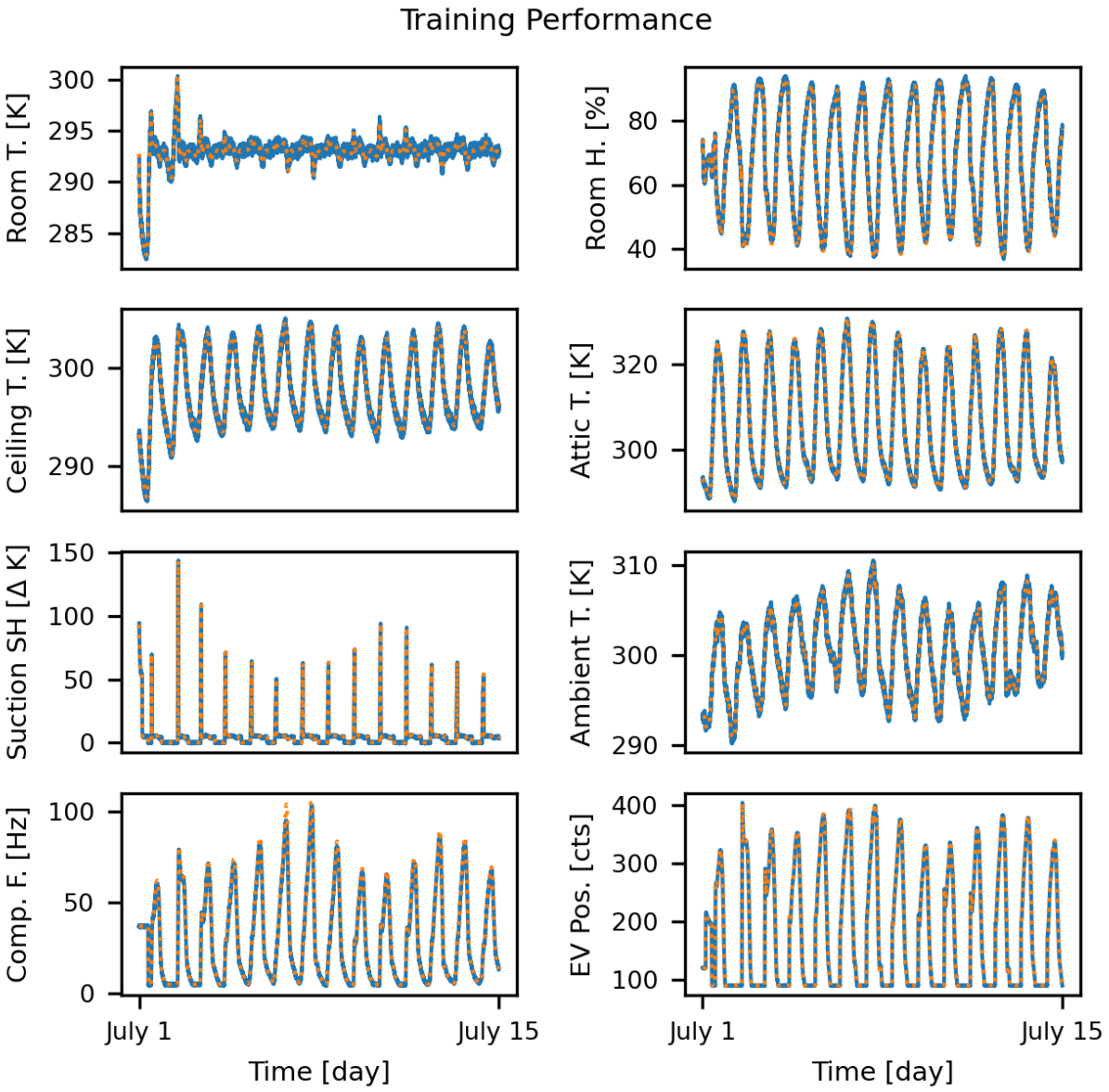

Figure 7: Performance of proposed algorithm on training data.

rather than the larger and more complex underlying dynamics. We demonstrated that the use of Gaussian processes as meta-models for the calibration-cost function is impractical in higher dimensions, and proposed the use of sparse GP approximations to address this issue. We also demonstrated the accuracy and efficacy of our proposed approach on a Modelica model of a building conditioned by vapor-compression equipment with 17 tunable parameters: indeed, we show that the parameters we estimate adhere to the ASHRAE guidelines. We also perform further testing on our proposed method and demonstrate that the method is robust to the initial set of samples used to construct the initial GP model.

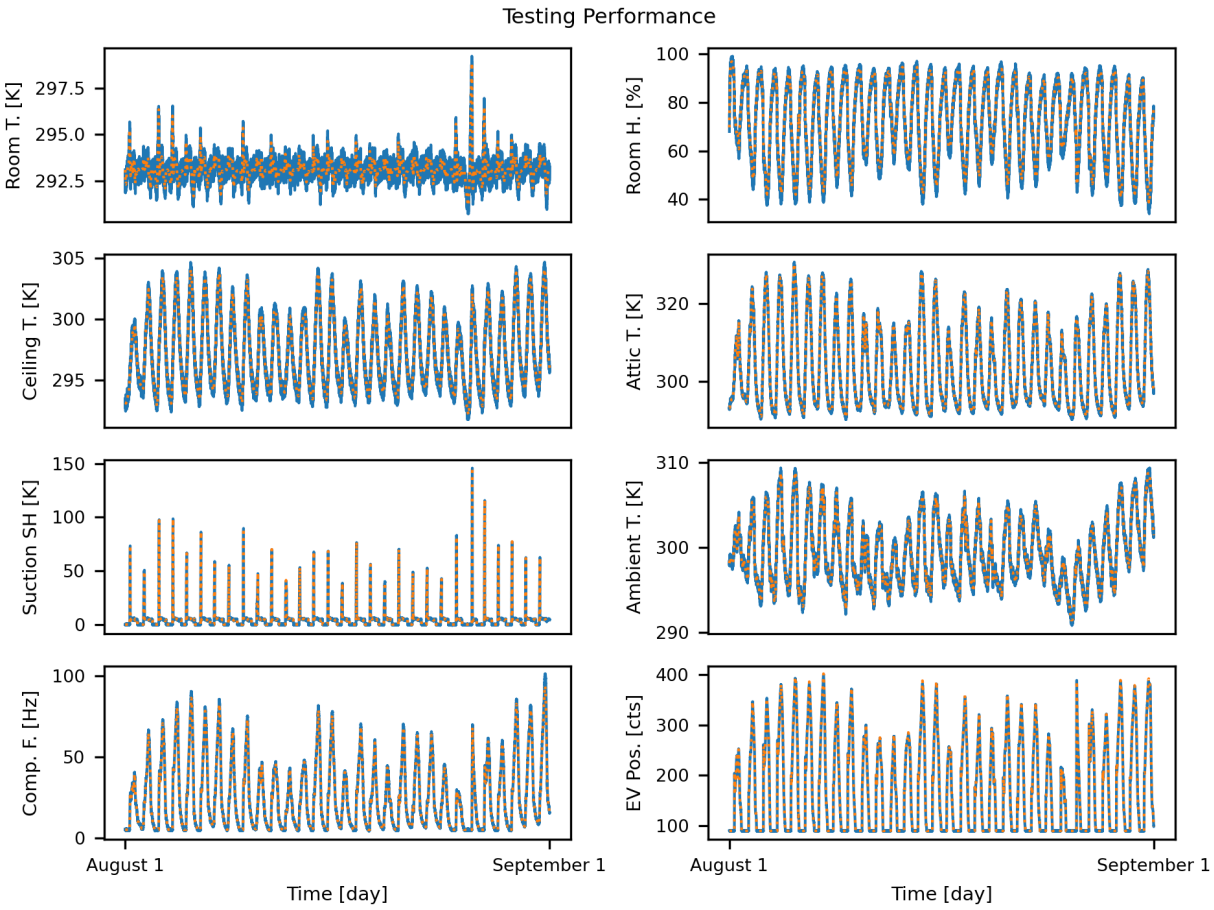

Figure 8: Performance of proposed algorithm on testing data.

Output	Variable Name	CVRMSE	|NMBE|	|MBE|
Room Temp.	`TRoom`	$< 0.1\%$	$< 0.1\%$	< 0.1 K
Room Humidity	`HumRoom`	$< 0.1\%$	$< 0.1\%$	2.2 %
Ceiling Temp.	`TCeiling`	$< 0.1\%$	$< 0.1\%$	< 0.1 K
Attic Temp.	`TAttic`	$< 0.1\%$	$< 0.1\%$	< 0.1K
Suction Superheat	`SHSuction`	$< 0.1\%$	$< 0.1\%$	0.12 K
Ambient Temp.	`TAmbient`	$< 0.1\%$	$< 0.1\%$	< 0.1 K
Compressor Freq.	`CF`	0.7%	$< 0.1\%$	< 0.1 Hz
Exp. Valve Pos.	`EEV`	$< 0.1\%$	$< 0.1\%$	< 0.1 count
Power ($\times$)	`Power`	6.87%	2.56	0.03 kW

Table 2: Output calibration performance metrics. $\times$ = not used for calibration.

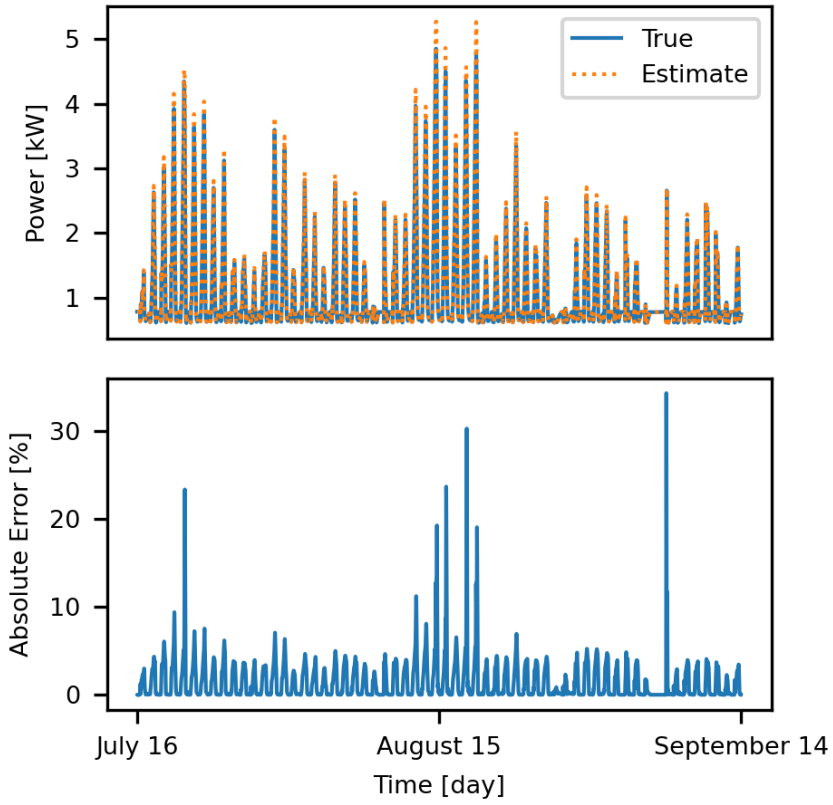

Figure 9: Comparison of power data and calibrated model power outputs.

The case study undertaken in this paper does reflect an idealized scenario in which the structure of the Modelica model is identical to the system that generates the data, so that the theoretical optimal error without noise could be zero. Because this is rarely true in practice, model calibration using real-world building data is more challenging because building models constructed in software will always have a non-zero 'sim-to-real' gap. However, the current work has demonstrated encouraging results, and the SGP-BO method described in this paper is robust to reasonable levels of measurement noise due to its probabilistic nature. For the scenario when the data has been corrupted by measurement noise or there is a small amount of model uncertainty, the cost function cannot be driven to zero, but SGP-BO still enables robust and efficient minimization with much fewer model simulations than existing calibration methods. As the level of this noise increases, the performance of SGP-BO or any other calibration method will degrade as the amount of useful information decreases. In other scenarios in which there are significant differences between the structures of the true physical system and the model, adaptations to the model structure are needed, either by redesigning sub-components of the simulation model or by appending to the existing model directly using data-driven methods, e.g., [44]. The use of SGP-BO for parameter estimation in simulation models

while concurrently optimizing hyperparameters for learning data-driven models for residual dynamics has not been studied in the current paper, but opens an interesting avenue for future research.

References

[1] D. Kim, et al., System identification for building thermal systems under the presence of unmeasured disturbances in closed loop operation, Building and Environment 107 (2016) 169–180.

[2] Z. Guo, et al., Identification of aggregate building thermal dynamic model and unmeasured internal heat load from data, in: Proc. Conf. on Decision and Control (CDC), IEEE, 2019, pp. 2958–2963.

[3] A. Garrett, J. New, Scalable tuning of building models to hourly data, Energy 84 (2015) 493–502.

[4] S. Asadi, et al., Building energy model calibration using automated optimization-based algorithm, Energy and Buildings 198 (2019) 106–114.

[5] S. A. Bortoff, C. R. Laughman, An Extended Luenberger Observer for HVAC Application using FMI, in: Modelica, 2019, pp. 157–015.

[6] M. Alam, et al., Applying extended Kalman filters to adaptive thermal modelling in homes, Advances in Building Energy Research 12 (1) (2018) 48–65.

[7] K. Bamdad, M. E. Cholette, J. Bell, Building energy optimization using surrogate model and active sampling, Journal of Building Performance Simulation 13 (6) (2020) 760–776.

[8] Y. Heo, et al., Scalable methodology for large scale building energy improvement: Relevance of calibration in model-based retrofit analysis, Building and Environment 87 (2015) 342–350.

[9] W. Tian, S. Yang, Z. Li, S. Wei, W. Pan, Y. Liu, Identifying informative energy data in bayesian calibration of building energy models, Energy and Buildings 119 (2016) 363–376.

[10] J. Berger, et al., Bayesian inference for estimating thermal properties of a historic building wall, Building and Environment 106 (2016) 327–339.

[11] Q. Li, G. Augenbroe, 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 (2016) 194–202.

[12] A. Chong, K. P. Lam, A comparison of MCMC algorithms for the bayesian calibration of building energy models, in: Proc. of the 15th IBPSA Building Simulation Conference, Vol. 4, 2017.

[13] K. Menberg, Y. Heo, R. Choudhary, Efficiency and Reliability of Bayesian Calibration of Energy Supply System Models, Proc. of the 15th IBPSA Building Simulation Conference (1) (2017) 1212–1221.

[14] A. Chong, W. Xu, S. Chao, N.-T. Ngo, Continuous-time bayesian calibration of energy models using BIM and energy data, Energy and Buildings 194 (2019) 177–190.

[15] H. Lim, Z. J. Zhai, Comprehensive evaluation of the influence of meta-models on Bayesian calibration, Energy and Buildings 155 (2017) 66–75.

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

[17] M. Quiroz, R. Kohn, M. Villani, M.-N. Tran, Speeding up MCMC by efficient data subsampling, Journal of the American Statistical Association 114 (526) (2018) 831–843.

[18] R. Zhang, et al., Optimal selection of building components using sequential design via statistical surrogate models, in: Proc. of the 13th IBPSA Building Simulation Conference, 2013, pp. 2584–2592.

[19] E. Gengembre, et al., A Kriging constrained efficient global optimization approach applied to low-energy building design problems, Inverse Problems in Science and Engineering 20 (7) (2012) 1101–1114.

[20] E. Tresidder, Y. Zhang, A. I. Forrester, Optimisation of low-energy building design using surrogate models, in: Proc. of Building Simulation, 2011, pp. 1012–1016.

[21] T. Østergård, R. L. Jensen, S. E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Applied Energy 211 (2018) 89–103.

[22] C. K. Williams, C. E. Rasmussen, Gaussian Processes For Machine Learning, Vol. 2, MIT press Cambridge, MA, 2006.

[23] J. Quiñonero Candela, C. E. Rasmussen, A unifying view of sparse approximate Gaussian process regression, JMLR 6 (Dec) (2005) 1939–1959.

[24] M. Bauer, M. van der Wilk, C. E. Rasmussen, Understanding probabilistic sparse gaussian process approximations, NeurIPS 29 (2016) 1533–1541.

[25] Y. Heo, et al., Evaluation of calibration efficacy under different levels of uncertainty, Journal of Building Performance Simulation 8 (3) (2015) 135–144.

[26] Z. Guo, et al., Aggregation and data driven identification of building thermal dynamic model and unmeasured disturbance, Energy and Buildings 231 (2021) 110500.

[27] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, NeurIPS 25 (2012) 2951–2959.

[28] T. D. Bui, J. Yan, R. E. Turner, A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation, JMLR 18 (1) (2017) 3649–3720.

[29] M. Titsias, Variational learning of inducing variables in sparse Gaussian processes, in: AISTATS, 2009, pp. 567–574.

[30] A. G. D. G. Matthews, J. Hensman, R. Turner, Z. Ghahramani, On sparse variational methods and the Kullback-Leibler divergence between stochastic processes, in: Artificial Intelligence and Statistics, PMLR, 2016, pp. 231–239.

[31] A. Chakrabarty, G. T. Buzzard, A. E. Rundell, Model-based design of experiments for cellular processes, Wiley Interdisciplinary Reviews: Systems Biology and Medicine 5 (2) (2013) 181–203.

[32] J. C. Helton, J. D. Johnson, C. J. Sallaberry, C. B. Storlie, Survey of sampling-based methods for uncertainty and sensitivity analysis, Reliability Engineering & System Safety 91 (10-11) (2006) 1175–1209.

[33] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of sobol indices for the gaussian process metamodel, Reliability Engineering & System Safety 94 (3) (2009) 742–751.

[34] N. Srinivas, A. Krause, S. M. Kakade, M. W. Seeger, Information-theoretic regret bounds for gaussian process optimization in the bandit setting, IEEE Transactions on Information Theory 58 (5) (2012) 3250–3265.

[35] T. Paananen, J. Piironen, M. R. Andersen, A. Vehtari, Variable selection for gaussian processes via sensitivity analysis of the posterior predictive distribution, in: Proc. 22nd Int. Conf. on Artificial Intelligence and Statistics (AISTATS), PMLR, 2019, pp. 1743–1752.

[36] H. Qiao, et al., Dynamic Characteristics of an R-410A Multi-split Variable Refrigerant Flow Air-conditioning System, in: 12th IEA Heat Pump Conf., 2017.

[37] M. Wetter, et al., Modelica buildings library, Journal of Building Performance Simulation 7 (4) (2014) 253–270.

[38] C. Laughman, et al., Modeling and control of radiant, convective, and ventilation systems for multizone residences, in: Proc. of Building Simulation 2019, 2019, pp. 1956–1963.

[39] Modelica Association, Functional Mockup Interface for Model Exchange and Co-Simulation, Version 2.0.1 (2019). URL `www.fmi-standard.org`

[40] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, A. G. Wilson, GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration, in: Advances in Neural Information Processing Systems, 2018.

[41] N. Fumo, M. J. Torres, K. Broomfield, A multiple regression approach for calibration of residential building energy models, Journal of Building Engineering 43 (2021) 102874. doi:https://doi.org/10.1016/j.jobe.2021.102874.

[42] ASHRAE, Guideline 14-2014, measurement of energy, demand, and water savings, American Society of Heating, Refrig-

erating, and Air Conditioning Engineers, Atlanta, Georgia (2014).

[43] G. R. Ruiz, C. F. Bandera, Validation of calibrated energy models: Common errors, Energies 10 (10) (2017) 1587.

[44] A. Chakrabarty, M. Benosman, Safe learning-based observers for unknown nonlinear systems using Bayesian optimization, Automatica 133 (2021) 109860. doi:https://doi.org/10.1016/j.automatica.2021.109860.

Appendix A. Correlation Analysis

In Fig. A.10, we also demonstrate the correlation amongst the measured outputs and parameters via a matrix of Pearson's correlation coefficients (only the triangular form is shown since the matrices are symmetric). From subplot (a), we deduce that the parameters obtained during the calibration task are not strongly cross-correlated. The reason is that, during calibration, the parameters are not sampled according a particular 'prior' distribution which could induce correlations by design, but rather chosen by SGP-BO iteratively, by exploiting the information contained in the surrogate model via the acquisition function. Conversely, from subplot (b), we observe that the measured outputs are highly (positively and negatively) correlated. In fact, other than suction superheat, all the other outputs are correlated: this is expected and corroborates the effect of the coupling between HVAC and building dynamics, since the measured outputs contain both HVAC and building quantities.

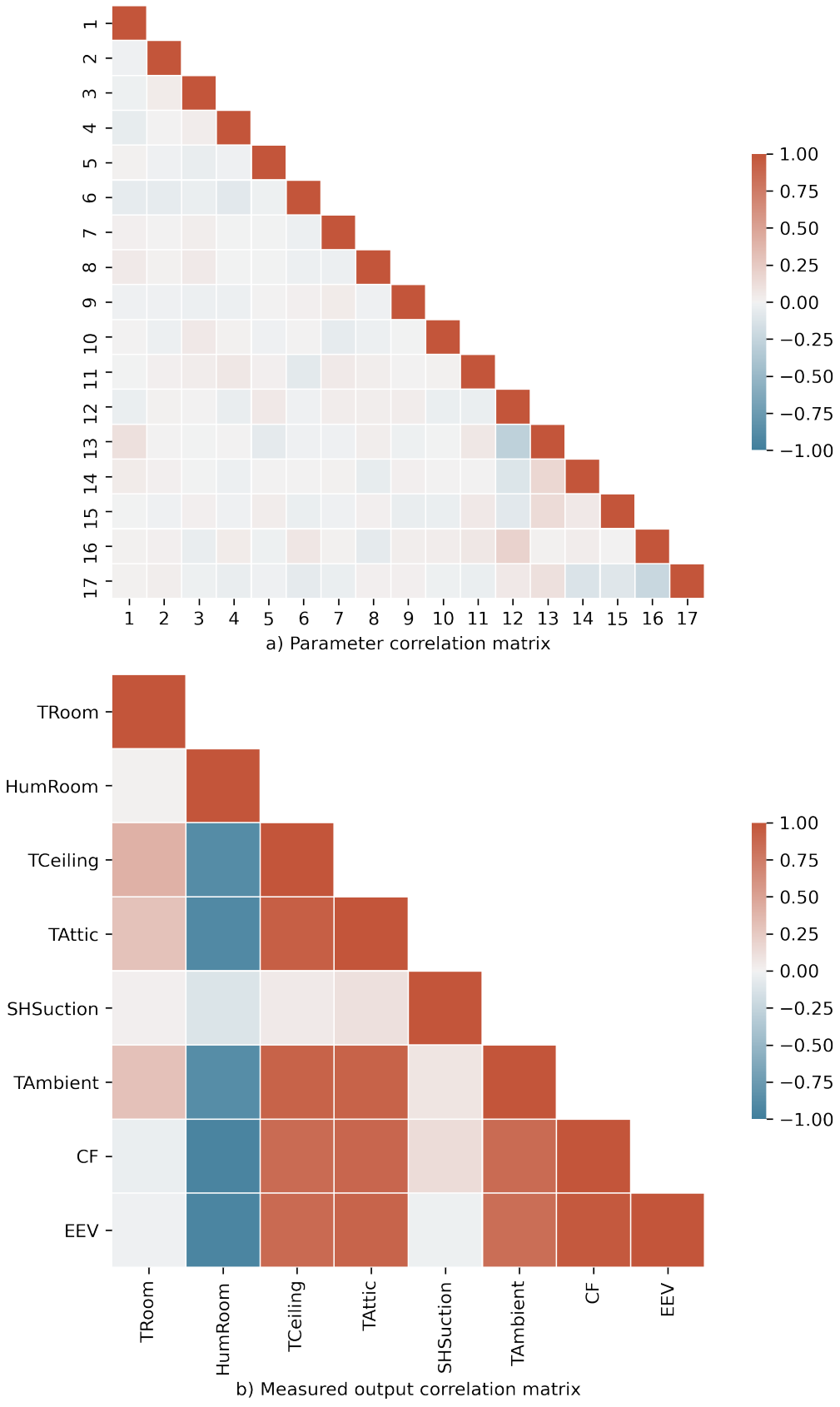

Figure A.10: Correlation matrices of the chosen outputs and parameters, obtained from data obtained during the calibration task. Variable descriptions are in Tables 1 and 2.