

EXPLAINABLE PARAMETER CALIBRATION VIA IMPORTANCE-DRIVEN SEQUENTIAL DESIGN WITH AN APPLICATION TO BUILDING ENERGY SYSTEMS

BY CHEOLJOON JEONG^{1,a} AND EUNSHIN BYON^{2,b}

¹*Department of Industrial Engineering, Clemson University, ajeong@clemson.edu*

²*Department of Industrial and Operations Engineering, University of Michigan, ebyon@umich.edu*

Parameter calibration seeks to estimate unobservable parameters in a computer model by aligning field observations with computer model outputs. In the building energy sector, a physics-based computer model is developed to analyze building energy use, given various weather conditions and operational scenarios. To obtain accurate simulations, it is necessary to calibrate model parameters required for preconfiguration. Among various techniques, Bayesian optimization stands out for its potential but faces some challenges when handling a large number of parameters. A possible remedy is to focus selectively on influential parameters, thereby simplifying a complex, high-dimensional task into a more tractable, lower-dimensional endeavor. We develop a new method that ranks parameter importance to effectively enable stochastic dimension reduction by utilizing the multi-armed bandit approach. By accounting for unequal importance among parameters, our approach generates accurate surrogate models tailored to the reduced dimension and guides an efficient exploration of the parameter search space in the Bayesian optimization procedure. The numerical studies and building energy simulation case study demonstrate that the proposed approach achieves a significant improvement in both calibration accuracy and efficiency. Moreover, it capably identifies the influential parameters and explains their impact on the computer model, providing valuable insights into understanding the system's dynamics.

1. Introduction. Advancements in numerical algorithms and computing capabilities have propelled computer models into a central role in the operational analysis and management of engineering systems. The computer models often incorporate complex mathematical functions derived from the first principles of physics. When applied appropriately, these computer models can simulate outputs that closely mirror the real-world physical dynamics. For instance, in the building sector, physics-based building energy models (BEMs) function as computerized tools that assist in analyzing the energy consumption patterns for the targeted buildings and thus help to enhance their energy efficiency. Specifically, these computer models have the capability to replicate a building's energy consumption by adjusting a range of factors such as heating, ventilation, and cooling (HVAC) settings, lighting configurations, plug loads, etc. One example of such a BEM simulator is EnergyPlus, which is developed by the U.S. Department of Energy (DOE)'s National Renewable Energy Laboratories (U.S. Department of Energy (2019)).

For simulation the computer model takes input variables, typically observable factors such as environmental and operational conditions, and generates corresponding outputs. Oftentimes, apart from input variables, additional parameters should be defined beforehand to run the simulation (Rutter et al. (2019)). For instance, in the BEM example, to simulate building energy processes, several parameters need to be set a priori. Table 1 lists a subset of parameters employed in EnergyPlus. The last three columns show the allowable ranges for each

Received September 2025; revised January 2026.

Key words and phrases. Bayesian optimization, efficient global optimization, multi-armed bandit, sequential design.

TABLE 1

List of BEM parameters and their ranges

Description	Unit	Default	Min	Max
Lighting level	W	590.37	300	700
Fraction radiant	—	0.32	0.26	0.38
Ventilation design flow rate	m ³ /s	0.001	0	1
Max supply air temperature	°C	48.44	30	70
Cooling supply air flow rate	m ³ /s	0.77	0	1
Heating supply air flow rate	m ³ /s	0.61	0	1
Rated air flow rate	m ³ /s	0.91	0.73	1
Heater thermal efficiency	—	0.77	0	1
Gross rated sensible heat ratio	—	0.74	0.6	0.8
Furnace nominal capacity	W	30,000	20,000	40,000
Burner efficiency	—	0.9	0.5	1
Gross rated cooling COP	—	2.95	2	5
Fan total efficiency	—	0.12	0	1
...	...	...	...	...

parameter, based on its physical requirement as well as the attributes of a building such as its type, location, and environmental factors. These parameters are designed to represent a target building’s real physical dynamics.

When these parameters are not directly observable, their values need to be estimated using either physical laws or data. In cases where physical laws are not available, they can be estimated using the data collected from the physical system. This process is referred to as *parameter calibration* in the literature (Kennedy and O’Hagan (2001), Jeong et al. (2023)). One of the predominant calibration approaches is Bayesian calibration (Kennedy and O’Hagan (2001), Higdon et al. (2004)). It has been employed for the BEM parameter calibration (Coakley, Raftery and Keane (2014), Chong and Menberg (2018)). Despite its widespread use, this method faces limitations with respect to scalability and computational efficiency. As an alternative, recent studies have explored nonlinear optimization techniques (Liu et al. (2022), Jeong et al. (2023)). These methods use a local search strategy, which can optimize parameters efficiently but with the caveat that the outcomes may be highly contingent upon the chosen starting points. We provide further discussion on the limitations of the Bayesian calibration and nonlinear optimization approaches in Section 3.

Bayesian optimization (BO) has recently been investigated as another promising approach, especially in building energy applications (Chakrabarty et al. (2021), Martinez-Viol et al. (2022), Zhan et al. (2022), Jeong and Byon (2024)). BO is a derivative-free optimization method to minimize an expensive-to-evaluate objective function that has no explicit expression. BO works by constructing a Gaussian process (GP) surrogate model for the unknown function and then designing the next query point through the optimization of an acquisition function. Nevertheless, BO faces challenges when dealing with a large number of calibration parameters (here “large” implies 20+ parameters). This is due to difficulties in accurately modeling the surrogate with GPs given a limited set of design points. Additionally, optimizing the acquisition function becomes more complex as the number of parameters grows.

An effective approach in BO is to strategically select a subset of parameters that strongly influence the computer model output. This involves reducing the original high-dimensional problem to a lower-dimensional one, referred to as *dimension slicing* in this paper. We explore this strategy based on *parameter importance*. The rationale behind our approach is that in many practical scenarios, only a handful of parameters significantly impact the computer model output—a principle described as *effect sparsity* (Linkletter et al. (2006), Ba, Myers and

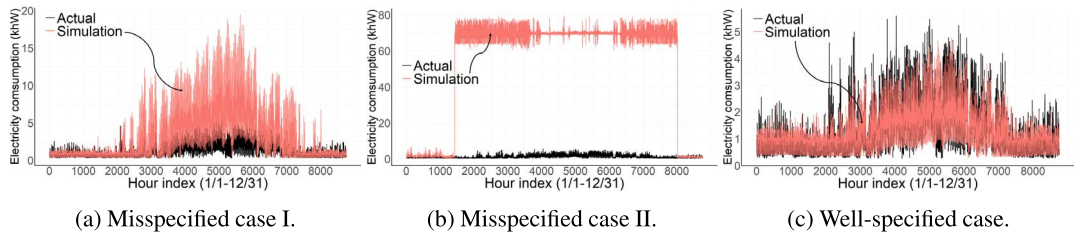


FIG. 1. Example of BEM simulation outputs.

Wang (2018), Salem et al. (2019), Park et al. (2025)). Additionally, the extent to which these important parameters affect the output can vary (Li, Ko and Byon (2021)), which is termed *effect imbalance*. These aspects are well documented and justified across a wide range of applications. When developing computer models, scientists and engineers often include many parameters to represent a wide range of physical phenomena and system configurations. For example, EnergyPlus incorporates dozens of parameters for HVAC systems, lighting, insulation, occupant behavior, etc. Yet under *specific* building types, operating and weather conditions, only a subset of these parameters could significantly affect outputs, while others remain nearly inactive or play only a secondary role. Thus, context-dependent sparsity and imbalance in their effects are both plausible and empirically supported in many simulation-based studies. Recognizing these patterns, we devise a new method that quantifies parameter importance to carry out the dimension slicing stochastically. This stochastic dimension slicing approach can potentially yield a more accurate surrogate model within the reduced dimensional space, consequently improving the effectiveness of BO’s sequential design process.

Our slicing approach also aids in achieving a more balanced trade-off between exploration and exploitation, especially when handling a large number of parameters. Traditional BO tends to choose design points close to the boundaries of the parameter space, where uncertainty is high; that is, more weight is given to exploration than to exploitation, in its early rounds (Binois and Wycoff (2022)). Although such exploration-focused search can be understandable, it could lead to model misspecification in some cases. To illustrate, one of the BEM parameters is the *cooling supply air flow rate*, which is deemed influential to the BEM output at the interior values. Another example parameter is the *fan total efficiency*, which is shown less influential in our case study. Though extreme values near 0 for these parameters are physically feasible, they rarely occur in practice and produce unrealistic simulation results as depicted in Figures 1(a) and (b). Such results could undesirably inflate the variance of the GP surrogate model used in BO. On the other hand, by implementing our slicing method within a reduced dimensional space, we naturally lower the uncertainty near the parameter space’s boundaries and avoid selecting unimportant parameters that might cause model misspecification. Consequently, our approach tends to encourage the selection of more moderate and realistic design points, generating reasonable simulation outcomes (see Figure 1(c)).

This study strategically reduces dimensionality through parameter selection, harnessing imbalanced effects among parameters to facilitate effective exploration. To capitalize on the effect sparsity and imbalance, we devise a weight distribution that reflects the parameter impact in the multi-armed bandit (MAB) framework. Our approach ensures that more influential parameters are prioritized during the calibration process, instead of treating all parameters equally. To our best knowledge, this study is the first to explicitly determine the importance of each parameter and incorporate this assessment into every sequential design during the calibration process within the BO procedure. Additionally, the weight distribution allows us to quantify the importance of each parameter. The numerical examples and the BEM case study demonstrate the superiority of our approach in comparison to other alternatives.

The rest of this article is organized as follows. Section 2 introduces the motivating application and dataset used in this study. Section 3 reviews related studies. Section 4 formulates the problem, presents the new calibration approach, and provides a regret bound analysis. Sections 5 and 6, respectively, present the numerical examples and BEM case study. Section 7 provides concluding remarks.

2. Motivating application and dataset. Energy consumption in buildings continues to rise globally due to economic development and the increasing electrification of end uses, making efficient energy management essential for power system stability and sustainability (U. S. Energy Information Administration (2015), Xu et al. (2021), González-Torres et al. (2022)). In Texas, for instance, frequent and intensifying heatwaves have pushed electricity demand to critical levels, straining grid reliability, raising operational costs, and contributing to social issues such as heat-related health emergencies, economic losses from outages, and heightened vulnerability among low-income households.

To address these challenges, researchers and practitioners have turned to real-world, data-rich *testbeds* that enable the study of residential energy consumption, the testing of new demand response strategies, and the evaluation of retrofit measures. A notable example is the *Pecan Street Project* in Austin, Texas (Pecan Street Inc. (2025)), which was launched to advance energy research and accelerate energy solutions by establishing a *living laboratory*. Pecan Street Inc., created as a research and development organization for this purpose, has deployed high-resolution sensors in residential buildings, capturing circuit-level electricity use, solar generation, environmental conditions, and appliance-specific loads. This dataset has been widely used in both academic and industrial studies, spanning topics such as load forecasting, demand response program design, and other related applications (Feng et al. (2021), Dinh, Lee and Kim (2022), Liu et al. (2023), Ramos et al. (2023), Chen et al. (2024)).

For the development of such new demand response programs and operational designs, it is essential to understand how different choices will perform before actual implementation. BEMs provide a powerful tool for simulating the impact of such interventions under varied conditions. By enabling controlled experiments in simulation, BEMs help quantify the effects of proposed policies, identify potential trade-offs, and reduce implementation risks.

However, the accuracy of any BEM-based analysis critically depends on how well the model represents the actual characteristics of each building, which are governed by calibration parameters (see Table 1). To properly utilize the BEM, these parameters must be calibrated to match simulated energy performance with measured data. Figure 2 illustrates the specific BEM calibration procedure studied in this paper.

In this study we leverage the *Pecan Street* dataset (Pecan Street Inc. (2025)) as a real-world source. Specifically, we focus on one residential building in the dataset to demonstrate our proposed methodology. The data span from January 1, 2014 to December 31, 2014, with an

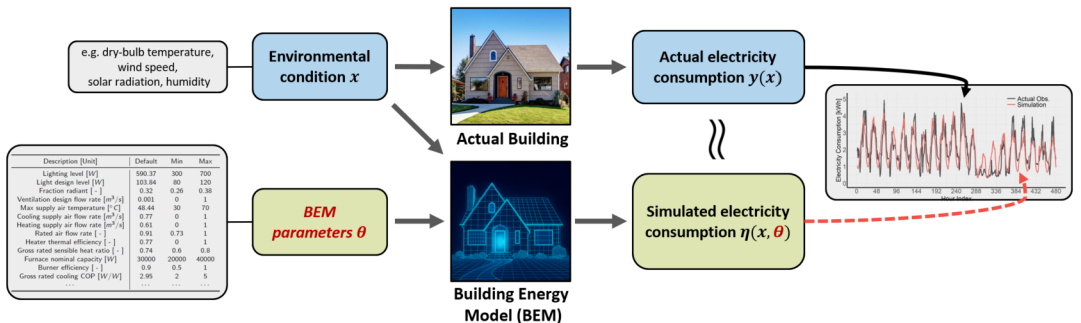


FIG. 2. Parameter Calibration for BEM.

hourly resolution; they include electricity consumption, environmental measurements, and metadata describing the building's geometry and characteristics.

However, applying BO with a large number of parameters often leads to excessive exploration along the boundaries of the parameter space. This, in turn, frequently results in model misspecification. Figures 1(a) and (b) illustrate this issue: the simulated BEM output deviates substantially from the measured energy consumption. Such discrepancies highlight the need for effective calibration methods that mitigate overexploration on the boundaries and that achieve a better balance between exploration and exploitation. To our knowledge, the model misspecification problem has not been accounted for in the BEM calibration literature.

3. Literature review. Bayesian calibration has been predominantly adopted for parameter calibration in the literature (Kennedy and O'Hagan (2001), Higdon et al. (2004), Chong and Menberg (2018)). Despite its widespread adoption, one of its challenges is scalability with respect to the number of parameters and data size. Typically, this method is limited to calibrating only a small handful of parameters (often up to five) because the computational complexity increases significantly with the number of hyperparameters in the covariance function. Further, due to substantial computational demands, it is generally applied to data with lower resolution. For instance, monthly (Li, Augenbroe and Brown (2016), Sokol, Davila and Reinhart (2017)) or annual (Booth, Choudhary and Spiegelhalter (2013)) data instead of hourly data for BEM calibration has been used. This limitation could potentially result in the loss of critical fine-grained information.

Gradient-based methodologies have been recognized for enhancing scalability in recent research (Liu et al. (2022), Jeong et al. (2023)). Nonetheless, they also bring along substantial computational demands, since they require a large number of simulation runs to approximate gradients and tune step sizes when dealing with black box computer models. The trust-region approach circumvents the need for direct gradient estimation by utilizing a local surrogate model to propose the next candidate point (Jain, Shashaani and Byon (2023, 2025)). Despite this advantage, their performance highly depends on the choice of the initial starting value.

Recently, BO has been gaining attention as an effective alternative for parameter calibration in several applications (Sha, Ozbay and Ding (2020), Reiker et al. (2021), Yang, Ji and Deng (2021), Ma et al. (2022)). However, the efficiency of BO decreases with the increase in the number of parameters. To address, two primary dimensionality reduction techniques have been proposed. The first is the parameter (or variable) selection approach, which identifies the influential parameters that substantially affect the model outputs. The second technique involves generating new parameters through linear or nonlinear combinations of the original parameters, known as linear or nonlinear embeddings (Binois and Wycoff (2022)). In the subsequent discussion, we will specifically focus on the parameter selection approach, because it is relevant when the model output is influenced predominantly by a subset of parameters.

The parameter selection approach typically involves the estimation of the lengthscales of an anisotropic kernel in GP modeling, referred to as automatic relevance determination (ARD) (Rasmussen and Williams (2006)). A small lengthscale implies that small changes with respect to its parameter lead to significant changes in the output. Utilizing the ARD, Salem et al. (2019) introduced an algorithm that iteratively selects influential variables. Yuan and Ng (2020) also utilize the ARD property for parameter selection through a Bayesian method, using a spike-and-slab prior. However, the estimated lengthscale may not accurately represent the true importance, because ARD generally decides the importance based on local variation, rather than considering long-range global changes (Wycoff, Binois and Wild (2021)).

Winkel et al. (2021) adopted a global variable selection approach proposed by Linkletter et al. (2006) and further complemented with the local variable selection. This approach can

result in substantial computational overhead due to the Markov chain Monte Carlo (MCMC) procedure, which can challenge its practicality. Thus, despite its potential to explain the variable importance, its use may not be suitable when the variable (i.e., parameter) selection method is applied in every round within the BO framework. Li et al. (2017) incorporated dropout techniques into BO by choosing a random subset of variables at each iteration. This approach does not account for variable importance in the selection process. Another approach is the trust-region-based BO (Regis (2016), Eriksson et al. (2019), Diouane et al. (2023)) that constrains the search space to a region surrounding the current best solution at each iteration. However, its performance could be sensitive to the choice of initial design points.

Additionally, it should be noted that most calibration problems with a large number of parameters prescreen the parameters first either using the parameter selection method or following the recommendation from domain experts and then calibrate only the selected parameters (Spagnol, Le Riche and Da Veiga (2019), Calama-González et al. (2021), Chong and Menberg (2018)). This two-phase approach with the parameter selection and calibration processes may lead to poor calibration accuracy. Recently, the integration of parameter selection and calibration processes has been considered (Yuan and Ng (2020)), but their method requires many prior assumptions in the Bayesian framework, limiting the method's applicability across diverse applications. Additionally, when selecting parameters, their use of the ARD property can lead to inaccuracies in quantifying the importance of parameters, as discussed earlier.

4. Methodology.

4.1. Problem formulation. Let $\theta \in \Theta \subseteq \mathbb{R}^D$ denote a vector of calibration parameters, where the range of the k th parameter is $[a_k, b_k]$ with a_k and b_k known constants for $k = 1, \dots, D$. Without loss of generality, we consider Θ as the normalized parameter space hereafter, that is, $\Theta = [0, 1]^D$. Let $\mathbf{x} \in \mathcal{X}$ denote a vector of physically observable inputs, where $\mathcal{X}$ is a convex and compact subset of $\mathbb{R}^M$. Let $\mathbf{y}(\mathbf{x})$ denote a vector of noisy field observations of its true physical process $\zeta(\mathbf{x})$ at an input $\mathbf{x}$. Let $\eta(\mathbf{x}; \theta)$ denote a vector of outputs from a deterministic computer model at $(\mathbf{x}, \theta)$, simulating $\zeta(\mathbf{x})$. In building energy simulation, $\mathbf{x}$ can be various weather conditions such as dry-bulb temperature, wind speed, solar radiation, and relative humidity for a building's surrounding area, $\mathbf{y}(\mathbf{x})$ values of measured energy consumption in the building, and $\eta(\mathbf{x}; \theta)$ values of energy consumption simulated by the BEM. Note that both $\mathbf{y}(\mathbf{x})$ and $\eta(\mathbf{x}; \theta)$ can be either scalar- or vector-valued responses, depending on the number of energy types in analysis, such as electricity, natural gas, steam energy, etc. In this study we consider the single-output electricity consumption due to limited data availability, but our proposed method can be easily applied to the multioutput setting.

To make $\eta(\mathbf{x}; \theta)$ resemble $\zeta(\mathbf{x})$, the unknown θ should be calibrated using operational data. Our goal is to learn the parameters θ^* that minimizes the expected squared L_2 distance between $\zeta(\mathbf{x})$ and $\eta(\mathbf{x}; \theta)$, that is,

$$(1) \quad \theta^* = \arg \min_{\theta \in \Theta} E_{\mathbf{x} \in \mathcal{X}} \|\zeta(\mathbf{x}) - \eta(\mathbf{x}; \theta)\|_{L_2}^2,$$

where $\|\cdot\|_{L_2}$ is the L_2 norm. One can use other types of loss instead of the squared L_2 loss.

In practice, the optimal parameters can be estimated by minimizing the empirical loss function $f(\theta)$, where the integral for the squared L_2 distance is replaced by its sample average from N_x observations and each $\zeta(\mathbf{x})$ by its noisy observation $\mathbf{y}(\mathbf{x})$. Thus, we formulate the parameter calibration problem to find the estimate $\hat{\theta}$ as

$$(2) \quad \hat{\theta} = \arg \min_{\theta \in \Theta} f(\theta),$$

where $f(\boldsymbol{\theta})$ denotes the empirical loss, defined as

$$(3) \quad f(\boldsymbol{\theta}) := \frac{1}{N_x} \sum_{i=1}^{N_x} \|\mathbf{y}(\mathbf{x}_i) - \boldsymbol{\eta}(\mathbf{x}_i; \boldsymbol{\theta})\|_2^2,$$

with $\|\cdot\|_2$ denoting the ℓ_2 or Euclidean norm.

In the literature the predominant calibration methods rely primarily on nonsequential sampling approaches, such as Bayesian calibration (Kennedy and O'Hagan (2001), Higdon et al. (2004)). These methods construct emulators for field observations, simulation outputs, or both at *predesigned* parameter values, particularly in limited data settings, whereas our BO-based calibration method directly constructs a surrogate of the empirical loss function with an initial dataset and *sequentially* designs the next parameter to evaluate its loss *on the fly* during the calibration process. Our sequential approach offers an advantage, as it generates more informative data from the computer model with the assistance of a continuously updated surrogate throughout the calibration process.

Since $\boldsymbol{\eta}(\mathbf{x}; \boldsymbol{\theta})$ in (3) does not take a closed form when the simulator is a black box computer model, BO can be an effective tool for solving (2). However, using BO blindly comes with various issues when handling a large number of parameters, say, 20+. First, standard BO does not prioritize parameters based on their importance, thereby treating all parameters with equal weight. Second, BO tends to prioritize exploration, particularly in the initial phase. While exploration is a critical component of the process to learn the response surface, excessive exploration often chooses design points at the boundaries of the parameter space when uncertainties are typically higher than the interior. This can potentially cause simulation runs with misspecified model configurations, as depicted in Figures 1(a) and (b).

We devise a new sequential design strategy that builds a GP surrogate using a subset of parameters in each BO round. These parameters are probabilistically selected based on their *importance measure*. This importance metric not only directs parameter selection but also quantifies and explains the degree of each parameter's importance. Thus, we refer to the proposed approach as “explainable sliced calibration” (briefly, Ex-Sliced-Calib).

4.2. Importance quantification via weight distribution. To assess the impact of parameters on the loss function $f(\boldsymbol{\theta})$ in (3), we first employ the global sensitivity analysis (GSA) approach (Iooss and Lemaître (2015)). We perform GSA based on a nonparametric surrogate $\hat{f}(\cdot)$, such as a kriging model. With all available data $\mathcal{D}_{n-1} = \{\boldsymbol{\theta}_{1:n-1}, f(\boldsymbol{\theta}_{1:n-1})\}$ obtained up to the $(n-1)$ th round, where $\boldsymbol{\theta}_{1:n-1} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}\}$ and $f(\boldsymbol{\theta}_{1:n-1}) = \{f(\boldsymbol{\theta}_1), \dots, f(\boldsymbol{\theta}_{n-1})\}$, the kriging model $\hat{f}(\boldsymbol{\theta})$ is a GP $Z_n \sim \text{GP}(\mu_{n-1}, K_{n-1})$, where μ_{n-1} and K_{n-1} are the mean vector and covariance matrix, respectively. For the kernel function, we use an anisotropic (or ARD) kernel. Specifically, we employ the anisotropic Matérn 5/2 kernel. This choice allows each dimension of the calibration parameters to have its own length-scale, which determines how sensitive the surrogate model is to variations along that specific direction. More details on the rationale for using an anisotropic instead of an isotropic kernel are given in the Supplementary Material (Jeong and Byon (2026)).

Treating $\boldsymbol{\Theta} \in [0, 1]^D$ as random variables, we conduct GSA using the following total variance decomposition (Sobol (2001)) based on a GP $Z_n(\boldsymbol{\theta})$, to obtain the total sensitivity index $T_{n,k}$ for the k th parameter θ_k at the n th round,

$$(4) \quad \text{Var}(Z_n(\boldsymbol{\Theta})) = \sum_{k \in \mathcal{K}} V_k + \sum_{k < k' \in \mathcal{K}} V_{k,k'} + \dots + V_{\mathcal{K}},$$

where $V_k = \text{Var}_{\boldsymbol{\Theta}_k}(E_{\boldsymbol{\Theta}_{-k}}[Z_n(\boldsymbol{\Theta})|\boldsymbol{\Theta}_k])$ is the contribution in total variance of each parameter θ_k for all k , $V_{k,k'} = \text{Var}_{\boldsymbol{\Theta}_{kk'}}(E_{\boldsymbol{\Theta}_{-kk'}}[Z_n(\boldsymbol{\Theta})|\boldsymbol{\Theta}_k, \boldsymbol{\Theta}_{k'}]) - V_k - V_{k'}$ is the portion of the total variance due to the interaction between θ_k and $\theta_{k'}$ for all $k < k'$, and $V_{\mathcal{K}}$ is the variance

contribution from the interaction among all of the parameters in $\mathcal{K} = \{1, \dots, D\}$. Here Θ_k is a random variable of the k th parameter, $\Theta_{kk'}$ is a random vector described by a joint distribution of Θ_k and $\Theta_{k'}$, Θ_{-k} is a random vector associated with a joint distribution of $\Theta_{\mathcal{K}}$, except Θ_k , and $\Theta_{-kk'}$ is one for $\Theta_{\mathcal{K}}$, except both Θ_k and $\Theta_{k'}$.

The variance decomposition is useful to quantify the variability of the response $Z_n(\Theta)$ with respect to each parameter's change (Gramacy (2020)). To measure the overall influence of the k th parameter, *total sensitivity indices* (Homma and Saltelli (1996)) can be defined as

$$(5) \quad T_{n,k} = 1 - \frac{\text{Var}_{\Theta_{-k}}(E_{\Theta_k}[Z_n(\Theta)|\Theta_{-k}])}{\text{Var}(Z_n(\Theta))} = \frac{E_{\Theta_{-k}}[\text{Var}_{\Theta_k}(Z_n(\Theta)|\Theta_{-k})]}{\text{Var}(Z_n(\Theta))} \quad \forall k \in \mathcal{K},$$

which quantifies the *total effect*, including all the first- and higher-order effects of θ_k , on the variance of the response $Z_n(\Theta)$.

To compute $T_{n,k}$, we use the extended Fourier amplitude sensitivity test (FAST) (Saltelli, Tarantola and Chan (1999)), recognized for its computational efficiency. Extended FAST generates input samples along a periodic curve $\vartheta_k(t) = G_k(\sin(\omega_k t))$, where $G_k(u) = 1/2 + (1/\pi) \arcsin(u)$ represents the k th transformation function, with each input associated with a frequency ω_k for each k . By varying t , the multidimensional input space is explored. These periodically sampled values are then propagated through the surrogate model by computing the Fourier transform for the model output $Z_n(\vartheta(t))$, where $\vartheta(t) = (\vartheta_1(t), \dots, \vartheta_D(t))$. If the model output is highly sensitive to variations in the k th parameter, the periodic function $Z_n(\vartheta_k(t))$ associated with ω_k will exhibit a large amplitude.

To quantify the importance of each parameter, we use the sensitivity index for each parameter obtained by extended FAST and introduce a weight distribution refined as new data points become available. Specifically, we utilize the MAB framework for dimension slicing, where each parameter is treated as an arm. Selecting a parameter is thus analogous to pulling the corresponding arm within the MAB context. When a subset of parameters is chosen simultaneously, it corresponds to pulling a combination of arms, an approach known as combinatorial MAB (Cesa-Bianchi and Lugosi (2012)). In each iteration, we select a subset of parameters that appear to have a significant impact on the loss in (3) by employing Thompson sampling (TS) (Russo et al. (2018)). TS facilitates probabilistic parameter selection, with the aim of efficiently navigating the parameter space and balancing exploration and exploitation.

Specifically, we treat $T_{n,k}$ as a new data point for the k th parameter's weight w_k , allowing it to play a role as a *reward* when refining its weight distribution. Here each parameter's weight is set to follow a Gaussian distribution with unknown mean λ_k and known precision $\tau_0 = 1/\sigma_0^2$. One can consider various distributions for priors if a particular belief is available. Otherwise, we can use the Gaussian distribution due to its conjugacy properties as

$$(6) \quad \pi(\lambda_k|w_k) \propto \pi(w_k|\lambda_k)\pi(\lambda_k) \quad \forall k,$$

with $\pi(w_k|\lambda_k) = \mathcal{N}(\lambda_k, \tau_0^{-1})$ and $\pi(\lambda_k) = \mathcal{N}(\nu_k, \tau_k^{-1})$, where ν_k and τ_k , respectively, denote location and precision parameters. The posterior is also Gaussian, and the model parameters ν_k and τ_k can be updated with the following closed-form expressions, treating $T_{n,k}$ as a new data point (here we put a subscript “ n ” to represent the round):

$$(7) \quad (\nu_{n,k}, \tau_{n,k}) \leftarrow \left(\frac{\tau_{n-1,k}\nu_{n-1,k} + \tau_0 T_{n,k}}{\tau_{n-1,k} + \tau_0}, \tau_{n-1,k} + \tau_0 \right) \quad \forall k.$$

When implementing this Bayesian update rule, we fix σ_0 for $\tau_0 = 1/\sigma_0^2$ to allow the weight distributions to adequately disperse, facilitating sufficient exploration of the parameter space.

Subsequently, borrowing the idea of bandits with knapsacks (Badanidiyuru, Kleinberg and Slivkins (2018)), we choose parameters of size d ($d < D$), referred to as the trial dimension. Specifically, we draw $\lambda_{n,k}$'s from the updated posterior distribution $\pi(\lambda_k|w_k)$ and sample

$w_{n,k}$'s from the k th weight distribution $\pi(w_k|\lambda_{n,k})$ for all k . In the implementation, we use the truncated Gaussian distribution for $\pi(w_k|\lambda_{n,k})$, which ranges from 0 to 1, ensuring that the drawn weights do not exceed this range. This means that the weights we draw from $\pi(w_k|\lambda_{n,k})$ cannot take a negative value or an excessively large positive value. We then select the d parameters with the largest weights among the D parameters.

The reason we conduct TS with the Bayesian update is that sensitivity indices may be unstable due to the potentially inaccurate GP surrogate. That is, we do not know the GP quality a priori. Additionally, even with the same data points, we may obtain different surrogates (or hyperparameter estimates), especially in high dimensions, due to an insufficient number of data points and the nonconvex nature of the likelihood functions that we maximize. By employing TS, we aim to enhance the robustness of our parameter selection process, mitigating the impact of any inaccuracies in the GP model and ensuring more reliable outcomes.

The weight distributions help to quantify the parameter influence. GSA would yield large values of $T_{n,k}$'s for important parameters, which increase ν_k and λ_k . Conversely, if certain parameters have minimal or no impact, the respective values of ν_k and λ_k would be small, resulting in a decrease in the probability of being selected. This characteristic allows us to discern which parameters receive greater emphasis during the calibration process. On the other hand, our MAB-based probabilistic selection approach provides a safeguard, ensuring that seemingly unimportant parameters are not completely overlooked during the calibration process, thus maintaining a comprehensive exploration of the parameter space.

4.3. Sliced sequential design. To leverage the effect sparsity and effect imbalance, we slice the parameter space into two subspaces: one with d selected parameters that are likely to be more influential and another with the remaining $D - d$ parameters. The idea of sliced design is to construct a GP in the lower-dimensional space associated with d selected parameters only and design its portion of the next design. We call the space of selected parameters *major subspace* and the space of remaining parameters *minor subspace* (Salem et al. (2019)). Note that the major and minor subspaces could change over iterations, because parameters are chosen probabilistically according to their updated weight distributions.

In the n th round, let $\mathcal{K}_n^{\text{Maj}}$ denote a set of indices of size d that corresponds to the selected parameters among all indices $\mathcal{K} = \{1, \dots, D\}$ and $\mathcal{K}_n^{\text{Min}}$ for the remaining parameters. Let $\boldsymbol{\theta}_n^{\text{Maj}} = \{\theta_{n,k} : \forall k \in \mathcal{K}_n^{\text{Maj}}\} \in \boldsymbol{\Theta}_n^{\text{Maj}} := [0, 1]^d$ denote a set of selected parameters (called *major parameters*) and $\boldsymbol{\theta}_n^{\text{Min}} = \{\theta_{n,k} : \forall k \in \mathcal{K}_n^{\text{Min}}\} \in \boldsymbol{\Theta}_n^{\text{Min}} := [0, 1]^{D-d}$ a set of remaining parameters (called *minor parameters*). Collectively, we denote $\boldsymbol{\theta}_n = (\boldsymbol{\theta}_n^{\text{Maj}}, \boldsymbol{\theta}_n^{\text{Min}})$.

With the two sliced subspaces, we design the new query point by focusing more on the major parameter design in the lower-dimensional space rather than the original parameter space. In the sliced major subspace, we place a GP prior on $f(\boldsymbol{\theta}_{1:n-1})$ against $\boldsymbol{\theta}_{1:n-1}^{\text{Maj}}$ using $\mathcal{D}_{n-1}^{\text{Maj}} = \{\boldsymbol{\theta}_{1:n-1}^{\text{Maj}}, f(\boldsymbol{\theta}_{1:n-1})\}$, where $\boldsymbol{\theta}_{1:n-1}^{\text{Maj}}$ denotes the major portion of $\boldsymbol{\theta}_{1:n-1}$ and $f(\boldsymbol{\theta}_{1:n-1})$ are its loss function values obtained up to the $(n-1)$ th round. Then we have the predictive distribution $Z_n^{\text{Maj}} \sim \text{GP}(\mu_{n-1}^{\text{Maj}}, K_{n-1}^{\text{Maj}})$ where μ_{n-1}^{Maj} is the mean vector and K_{n-1}^{Maj} is the covariance matrix constructed from $\mathcal{D}_{n-1}^{\text{Maj}}$. Note that we incorporate a nugget effect into the diagonal entries of the base covariance matrix, which can be reestimated by maximum likelihood estimation at each round. The nugget term accounts not only for potential observational noise but also for the variability introduced by the noninjective nature of the projection, that is, cases where different parameter sets in the original space are mapped to the same point in the reduced space but yield distinct outputs. Further details about the nugget term are provided in the Supplementary Material (Jeong and Byon (2026)). With Z_n^{Maj} , we design the major portion of new query point $\boldsymbol{\theta}_n^{\text{Maj}}$ by optimizing an acquisition function.

Algorithm 1 Explainable Calibration via Sliced Sequential Design (Ex-Sliced-Calib)

-
- 1: **Input:** N_0 and $N_{\max}$ ($> N_0$)
 - 2: Evaluate f at N_0 points of θ , generated by the space-filling design (e.g., MaxPro), to get $\mathcal{D}_{N_0} = \{\theta_{1:N_0}, f(\theta_{1:N_0})\}$.
 - 3: Initialize the weight distributions $\pi(w_k | \lambda_{N_0,k})$, and prior hyperparameters $(v_{N_0,k}, \tau_{N_0,k})$, $\forall k$.
 - 4: **for** $n = N_0 + 1, \dots, N_{\max}$ **do**
 - 5: Construct a surrogate model $Z_n(\cdot)$ for $f(\cdot)$ using $\mathcal{D}_{n-1} = \{\theta_{1:n-1}, f(\theta_{1:n-1})\}$.
 - 6: Calculate sensitivity indices $T_{n,k}$ based on $Z_n(\cdot)$.
 - 7: Obtain posteriors $\pi(\lambda_k | w_k)$ by updating $(v_{n,k}, \tau_{n,k})$ via (6)–(7) with $T_{n,k}$, $\forall k$.
 - 8: **for** $k = 1, \dots, D$ **do**
 - 9: Sample $\lambda_{n,k} \sim \pi(\lambda_k | w_k)$.
 - 10: Sample $w_{n,k} \sim \pi(w_k | \lambda_{n,k})$.
 - 11: **end for**
 - 12: Select d parameters with the largest values among the D parameters.
 - 13: Construct a surrogate model $Z_n^{\text{Maj}}(\cdot)$ for $f^{\text{Maj}}(\cdot)$ with a major portion of samples $\mathcal{D}_{n-1}^{\text{Maj}} = \{\theta_{1:n-1}^{\text{Maj}}, f(\theta_{1:n-1}^{\text{Maj}})\}$.
 - 14: Design the major parameters θ_n^{Maj} by optimizing an acquisition function $\mathcal{A}(\theta^{\text{Maj}})$.
 - 15: Design the minor parameters θ_n^{Min} by employing a sequential space-filling design (e.g., sequential MaxPro) with probability p , or setting $\theta_n^{\text{Min}} = \theta_{\min}^{\text{Min}}$ with probability $1 - p$.
 - 16: Set the new query design θ_n by combining θ_n^{Maj} and θ_n^{Min} , that is, $\theta_n = (\theta_n^{\text{Maj}}, \theta_n^{\text{Min}})$.
 - 17: Evaluate f at θ_n and set $\mathcal{D}_n = \mathcal{D}_{n-1} \cup (\theta_n, f(\theta_n))$.
 - 18: **end for**
 - 19: **Output:** $(\hat{\theta}, f(\hat{\theta})) = (\theta_{\min}, f(\theta_{\min}))$ and $\pi(w_k | \lambda_{N_{\max},k})$, $\forall k$.
-

Next, we design the minor subspace by combining two approaches. First, we use the minor portion of the parameter design that provides the lowest loss function value among $f(\theta_{1:n-1})$, denoted by $\theta_{\min}^{\text{Min}}$ (Li et al. (2017)). This approach emphasizes exploitation by prioritizing the historically most promising area of the minor subspace. Second, we expand our search across the minor subspace with an additional space-filling design, facilitating more exploration to potentially uncover new regions of interest. This is because the surrogate for GSA might be inaccurate when the initial design does not sufficiently cover the parameter space. In such cases, important parameters could inadvertently be excluded from the major subspace. Thus, we allocate a small probability for additional space-filling to augment the existing design set. In this study we employ the sequential MaxPro design (Ba, Myers and Wang (2018)),

$$(8) \quad \theta_n^{\text{Min}} = \begin{cases} \text{sequential space-filling design} & \text{w.p. } p, \\ \theta_{\min}^{\text{Min}} & \text{otherwise.} \end{cases}$$

Finally, we design the new query point by combining the obtained major and minor parameter designs, that is, $\theta_n = (\theta_n^{\text{Maj}}, \theta_n^{\text{Min}})$. Then f is evaluated at θ_n by running the computer model, and $(\theta_n, f(\theta_n))$ is added to $\mathcal{D}_{n-1}$, forming $\mathcal{D}_n$. This procedure is repeated until convergence or the available budget $N_{\max}$ is reached. Algorithm 1 summarizes the procedure.

4.4. Regret bound analysis. We present a regret bound analysis that accounts for the characteristics of effect sparsity and imbalance. In our analysis we employ lower confidence bound (LCB) (Srinivas et al. (2010)) as the acquisition function, as it allows for tractable analysis and has been widely used to derive regret bounds in BO.

Let $\mathcal{K}^I$ denote a set of indices of size d_I that correspond to the influential parameters and $\mathcal{K}^{NI}$ a set of indices of size $D - d_I$ for the noninfluential parameters. We divide the parameters θ into two distinct groups of influential parameters, θ^I as $\theta^I = \{\theta_k : \forall k \in \mathcal{K}^I\}$, and noninfluential parameters, θ^{NI} as $\theta^{NI} = \{\theta_k : \forall k \in \mathcal{K}^{NI}\}$, so that we have $\theta = (\theta^I, \theta^{NI})$. To facilitate the theoretical analysis, we consider a simplified version of Ex-Sliced-Calib, summarized in Algorithm 1 in the Supplementary Material (Jeong and Byon (2026)), where the importance of each parameter is assumed to be known a priori (in the form of (A.1) and (A.2) in Assumption A.2 of the Supplementary Material (Jeong and Byon (2026))). It fixes the minor design point at θ_0^{NI} but chooses the major design point that optimizes the acquisition function.

We define the instantaneous regret $r_n = f(\theta_n) - f(\theta^*)$ at the n th round as we aim to find a minimizer θ^* . Then the cumulative regret R_N up to the N th round is calculated as $R_N = \sum_{n=1}^N r_n$, whose desirable asymptotic property is to have no asymptotic average regret, that is, $\lim_{N \rightarrow \infty} R_N/N = 0$. In this study we present an upper bound of R_N/N for the simplified Ex-Sliced-Calib. Here we present the main results. Other details, including assumptions and proofs, are provided in the Supplementary Material (Jeong and Byon (2026)).

THEOREM 4.1. *Suppose that Assumptions A.1–A.2 hold. Pick $\delta \in (0, 1)$, and define*

$$(9) \quad \begin{aligned} \beta_n = & 2 \log \left(\frac{2\pi^2 n^2}{3\delta} \right) + 2d_I \log \left[\alpha_{\max} D n^2 b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} \right] \\ & + 2(D - d_I) \log \left[\rho_{\max} D n^2 b \sqrt{\log \left(\frac{8Dc}{\delta} \right) + 1} \right], \end{aligned}$$

where $\alpha_{\max} = \max\{\alpha_k : \forall k \in \mathcal{K}^I\}$ and $\rho_{\max} = \max\{\rho_k : \forall k \in \mathcal{K}^{NI}\}$. With the simplified Ex-Sliced-Calib, we have the following upper bound for the averaged regret:

$$(10) \quad \frac{R_N}{N} \leq 2\sqrt{C_1 \frac{\beta_N \gamma_N}{N}} + \frac{\pi^2}{3N} + \rho_{\max} b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} (D - d_I),$$

with probability $\geq 1 - \delta$ for some constant C_1 , where γ_N is the maximum information gain with a finite set of N sampling points.

In Theorem 4.1 the information gain quantifies the reduction in uncertainty about f (Srinivas et al. (2010)) achieved by collecting a new data point or adding a new design point and obtaining the associated loss function value. With the popularly used kernel functions, such as the squared exponential kernel with the same lengthscales, γ_N is bounded above, as shown in Srinivas et al. (2010). Thus, we have $\beta_N \gamma_N / N \rightarrow 0$. The proof of Theorem 4.1 closely follows those in Srinivas et al. (2010) and Shen and Kingsford (2023). Compared to existing regret bounds in the literature, our bound introduces two key developments. First, it incorporates a decomposition of β_N over influential and noninfluential parameters. Second, it includes an additional ρ_k -related term to capture the residual influence of noninfluential parameters and explicitly highlight imbalance effects among parameters through ρ_k . These developments are new and particularly relevant for calibration tasks in which only a subset of parameters significantly affects the outcome.

Theorem 4.1 establishes an upper bound on the average regret when the calibration efforts focus exclusively on those parameters, under the assumption that the true influential parameters are known. This result provides important insights: the last term in equation (10) captures the cost associated with the noninfluential parameters, and it vanishes when $\rho_{\max}$ is zero. When $\rho_{\max} = 0$, meaning that all noninfluential parameters have no impact on the

computer model output or the loss function f , the simplified Ex-Sliced-Calib asymptotically achieves zero average regret with large N . Thus, if Ex-Sliced-Calib correctly identifies influential parameters, the regret bound in equation (10) also holds. Furthermore, Ex-Sliced-Calib can achieve zero average regret asymptotically when noninfluential parameters have no impact. Our numerical study in Section 5 demonstrates that Ex-Sliced-Calib correctly pinpoints influential parameters and quantifies their importance.

While we employ LCB in our regret bound analysis, we use the expected improvement (EI) (Moćkus (1975), Jones, Schonlau and Welch (1998)) acquisition function in our numerical and case studies, because EI consistently demonstrates superior empirical performance in terms of calibration accuracy. We are unable to derive a regret bound for EI, but we expect a similar or even tighter bound, given the outstanding results observed in the numerical study. For the empirical comparison of the acquisition functions, see the Supplementary Material (Jeong and Byon (2026)).

4.5. Uncertainty quantification. Taking effect sparsity and imbalance into consideration, we present a method for quantifying the uncertainty of the calibrated parameters using their asymptotic properties as maximum likelihood (ML) estimators. With the Ex-Sliced-Calib method, we can acquire the ML estimator $\hat{\theta}_{\text{ML}}$ for the calibration parameters if θ converges to θ^* under the normality assumption. This is due to the fact that maximizing the log likelihood when the observational error ϵ follows a normal distribution within the linear linkage model $\mathbf{y}(\mathbf{x}) = \eta(\mathbf{x}; \theta) + \epsilon$ is equivalent to minimizing the empirical loss function $f(\theta)$ (Jeong et al. (2023)). This allows us to consider the attractive theoretical traits for the calibrated parameters, including consistency and asymptotic normality. Since we consider only a subset of parameters that affect the model output, we focus on investigating the asymptotic properties of the ML estimator $\hat{\theta}_{\text{ML}}^I$ for the influential parameters only.

Once we finally identify the influential parameters and their ML estimator $\hat{\theta}_{\text{ML}}^I$, the consistency property dictates that $\hat{\theta}_{\text{ML}}^I$ converges in probability to the true influential parameter vector θ_0^I as $N_x \rightarrow \infty$, say, $\hat{\theta}_{\text{ML}}^I \xrightarrow{p} \theta_0^I$. Further, the asymptotic normality indicates that the estimator $\sqrt{N_x}(\hat{\theta}_{\text{ML}}^I - \theta_0^I)$ converges in distribution to a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0^I)^{-1})$ as $N_x \rightarrow \infty$, that is, $\sqrt{N_x}(\hat{\theta}_{\text{ML}}^I - \theta_0^I) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0^I)^{-1})$, where $\mathcal{I}(\theta_0^I)$ is an expected Fisher information matrix. Then we can use the ML estimator's asymptotic properties to construct the confidence intervals (CIs) of the influential parameters for their uncertainty quantification. The asymptotic 100(1 - α)% Wald CI for each component of $\hat{\theta}_{\text{ML}}^I$ as follows:

$$(11) \quad \hat{\theta}_{\text{ML},k}^I \pm z_{1-\alpha/2} \frac{1}{\sqrt{N_x}} \sqrt{\mathcal{I}_{kk}^{-1}(\hat{\theta}_{\text{ML}}^I)} \quad \forall k \in \mathcal{K}^I,$$

where $z_{1-\alpha/2}$ is a critical value of the normal distribution, and $\mathcal{I}_{kk}^{-1}(\cdot)$ denotes the k th diagonal entry of the inverse of the Fisher information matrix $\mathcal{I}(\cdot)$. The detailed method for estimating the Fisher information matrix in (11) is provided in the Supplementary Material (Jeong and Byon (2026)).

5. Numerical study. We demonstrate the effectiveness of the proposed calibration approach through numerical examples. The physical process is

$$(12) \quad y(x_i) = \exp(x_i/10) \sin(x_i) + \epsilon_i,$$

and the computer model is

$$(13) \quad \eta(x_i; \theta) = \exp(x_i/10) \sin(x_i) + \sqrt{\varphi(\theta_1, \dots, \theta_{d_I})},$$

for $\theta \in \mathbb{R}^D$, where d_I ($\ll D$) denotes the number of true influential parameters, $x_i \sim \text{Uniform}(0, 2\pi)$ represents an observable input value, and $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ is an observational error for all $i = 1, \dots, 200$. While we assume Gaussian noise, the proposed method remains relatively robust to moderate violations of this assumption as long as the underlying function is sufficiently smooth and the noise is not highly structured or adversarial. Further details and comparative results with non-Gaussian noise can be found in the Supplementary Material (Jeong and Byon (2026)). For $(\theta_1, \dots, \theta_{d_I}) \in [LB, UB]^{d_I}$, where LB and UB denote the lower and upper bounds of parameters, respectively, we modify the three well-known test functions for $\varphi(\cdot)$ in the computer model which have many local minima and are widely used in evaluating the optimization performance (Surjanovic and Bingham (2013)) as follows:

- Example I (Modified Schwefel)

$$(14) \quad \varphi(\theta_1, \dots, \theta_{d_I}) = 418.9829d_I - \sum_{k=1}^{d_I} \theta_k \sin(\sqrt{|\theta_k|}) + h(d_I, LB, UB, EV),$$

where $d_I = 2$, $LB = -500$, $UB = 500$, and $EV = 4,000$.

- Example II (Modified Ackley)

$$(15) \quad \varphi(\theta_1, \dots, \theta_{d_I}) = -c_1 \cdot \exp\left[-c_2 \sqrt{\frac{1}{d_I} \sum_{k=1}^{d_I} \theta_k^2}\right] - \exp\left[\frac{1}{d_I} \sum_{k=1}^{d_I} \cos(c_3 \theta_k)\right] + c_1 \\ + \exp(1) + h(d_I, LB, UB, EV),$$

where $d_I = 2$, $c_1 = 20$, $c_2 = 0.2$, $c_3 = 2\pi$, $LB = -5$, $UB = 10$, and $EV = 400$.

- Example III (Modified Hartmann3)

$$(16) \quad \varphi(\theta_1, \dots, \theta_{d_I}) = -\sum_{j=1}^4 c_j \exp\left[-\sum_{k=1}^3 A_{jk}(\theta_k - P_{jk})^2\right] + 3.86278 \\ + h(d_I, LB, UB, EV),$$

where $d_I = 3$, $\mathbf{c} = [1.0, 1.2, 3.0, 3.2]^T$, $\mathbf{A} = [3 \ 10 \ 30; 0.1 \ 10 \ 35; 3 \ 10 \ 30; 0.1 \ 10 \ 35]$, $\mathbf{P} = [3689 \ 1170 \ 2673; 4699 \ 4387 \ 7470; 1091 \ 8732 \ 5547; 381 \ 5743 \ 8828]$, $LB = 0$, $UB = 1$, $EV = 250,000$, and “3.86278” is the term to make the computer model perfect.

Note that we modify the three well-known functions by adding a boundary perturbation term $h(\cdot)$, which is designed to emulate model misspecification frequently observed in practical BEM implementations. This term $h(\cdot)$ takes a role in extreme values near the boundaries, mimicking how BEM simulations become unreliable at such points. These modifications enable us to evaluate how effectively Ex-Sliced-Calib handles model misspecification through informed exploration in the reduced parameter space. The term $h(\cdot)$ is expressed by

$$(17) \quad h(d_I, LB, UB, EV) = \sum_{l \in \mathcal{K}^I} \mathbb{1}\left(LB \leq \theta_l \leq \frac{24}{25}LB + \frac{1}{25}UB\right) \left[EV - \frac{\theta_l - LB}{UB - LB}\right] \\ + \sum_{l \in \mathcal{K}^I} \mathbb{1}\left(\frac{1}{25}LB + \frac{24}{25}UB \leq \theta_l \leq UB\right) \left[\theta_l - LB - \frac{24}{25}UB\right],$$

where $\mathbb{1}(\cdot)$ is an indicator function. The term including $\mathbb{1}(LB \leq \theta_l \leq \frac{24}{25}LB + \frac{1}{25}UB)$ represents the situation where the boundaries near the LB of the parameters produce unreasonably large values with the extreme values (EVs) and their extent linearly decreases from the boundaries to the interiors of the parameter space. The term including $\mathbb{1}(\frac{1}{25}LB + \frac{24}{25}UB \leq \theta_l \leq UB)$

TABLE 2
Problem settings in numerical examples

Example	Test function φ	Domain of θ	σ	d_I	D	N_0	$N_{\max}$
I	Modified Schwefel	$[-500, 500]^D$	0.1	2	30	150	500
II	Modified Ackley	$[-5, 10]^D$	0.1	2	30	150	500
III	Modified Hartmann3	$[0, 1]^D$	0.01	3	30	150	500

minimally inflates the outcomes around the boundaries near the UB and does not cause significantly unreasonable outcomes. More details, along with a performance comparison of the proposed method with and without incorporating $h(\cdot)$, are provided in the Supplementary Material (Jeong and Byon (2026)).

Table 2 summarizes the problem settings of numerical examples. Note that the experimental setup in Table 2 assumes that most parameter effects are sparse, capturing a key characteristic of our motivating application. However, we also conduct additional experiments in which the number of true influential parameters, d_I , is increased. The results and their implications are provided in the Supplementary Material (Jeong and Byon (2026)). All experiments were conducted on a 64-bit Windows OS with an Intel Xeon CPU E5-2697 and R version 4.2.2.

5.1. Quantifying parameter importance. One of the most salient features of Ex-Sliced-Calib is its explainability by quantifying the parameter importance through the weight distribution. Figure 3 shows the trajectories of each parameter's posterior mean of its weight distribution over the rounds. Clearly, the trajectories for important parameters, θ_1 and θ_2 for Examples I and II, and θ_2 and θ_3 for Example III, have higher posterior means. We note that even the weights of parameters deemed unimportant are kept positive, allowing for their occasional selection due to the probabilistic nature of weight distributions. This ensures that underrepresented parameters also receive consideration throughout the calibration process.

To check if Ex-Sliced-Calib captures the true influence correctly, we calculate total sensitivity indices without a surrogate using Monte Carlo methods, as described in Saltelli (2002). We treat these Sobol' indices as *true importance* of each parameter and compare the posterior mean obtained by Ex-Sliced-Calib at the maximum budget of $N_{\max} = 500$. Figure 4 shows that Ex-Sliced-Calib captures the true importance successfully. In Figure 3, θ_1 in Example III does not get much emphasis, compared to θ_2 and θ_3 . This is because θ_2 and θ_3 have much more impact on the computer model output than θ_1 with an order $\theta_3 > \theta_2 > \theta_1$ as can be seen in Figure 4(c). The true and estimated importance measures of all parameters can be found in the Supplementary Material (Jeong and Byon (2026)).

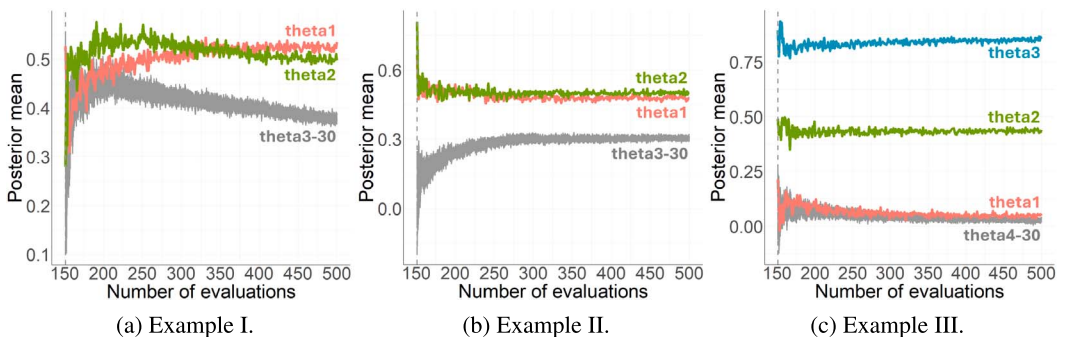


FIG. 3. Trajectories of posterior means for Ex-Sliced-Calib in the numerical study.

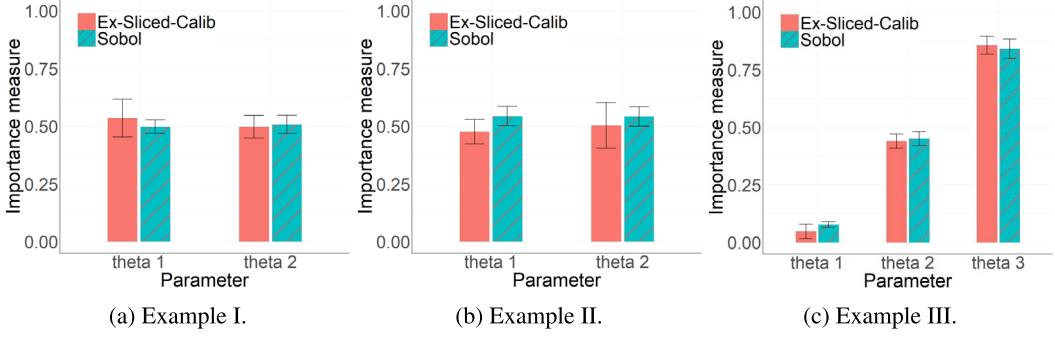


FIG. 4. Comparison of posterior means from Ex-Sliced-Calib with the true importance using Sobol' indices without constructing a GP surrogate in the numerical study.

5.2. *Comparison with other alternatives.* We consider three BO-based alternatives. The first and third ones below are proposed in the literature to handle the high-dimensional BO:

- *Random sliced design (Rand-Sliced-Calib)* selects a subset of parameters purely at random without considering their importance. It designs the next query point with the selected ones from the acquisition function optimization as well as the minor ones giving the lowest loss value until each round. This approach is called *BO using dropout* (Li et al. (2017)).
- *Full design (Full-Calib)* is the plain BO called efficient global optimization (EGO) (Jones, Schonlau and Welch (1998)). It designs the next query point with all parameters via the standard sequential design scheme without dimensionality reduction.
- *Trust-region-based BO (TR-Calib)* does not slice the parameter space into subspaces. Rather, it dynamically restricts the search space around each iteration's best point to choose the next point at every round as one of the remedies when D is large (Diouane et al. (2023)).

We first design initial points of size $N_0 = 5D$ using MaxPro (Joseph, Gul and Ba (2015)) due to its effective space-filling property on projections to lower-dimensional subspaces. Each method terminates when the number of rounds reaches its maximum budget of $N_{\max} = 500$. Each algorithm is run for 20 experiments with different random seeds, using the same seed across all competing algorithms. For both Ex-Sliced-Calib and Rand-Sliced-Calib, we set the trial dimension $d = 10$. Also, we set $p = 0.05$ and $\sigma_0 = 0.5$ for Ex-Sliced-Calib.

Table 3 reports the calibration results, and Figure 5 shows each method's trajectory of the mean value of the lowest MSEs over the rounds, with its ± 1 standard error band around the mean. Across all rounds, both Ex-Sliced-Calib and Rand-Sliced-Calib consistently outperform Full-Calib and TR-Calib. This notable improvement is due to their operation in the reduced dimension and designing the next query point in a low-dimensional subspace. Between the two, Ex-Sliced-Calib achieves lower MSEs, thanks to its emphasis on influential parameters. Also, Ex-Sliced-Calib generally yields smaller variance of MSEs than others, demonstrating its robustness. In Examples II and III, while parameter values obtained from Rand-Sliced-Calib are closer to the true ones than Ex-Sliced-Calib, its MSE is higher, possibly due to the large variance in Rand-Sliced-Calib.

Full-Calib exhibits a slower decrease in MSE across iterations, leading to inaccurate calibrated values with the limited budget. TR-Calib, on the other hand, demonstrates improved outcomes over Full-Calib, because it avoids exhaustive searches across the entire search space through the construction of an adaptive trust region. However, since it still operates within the original high-dimensional space, its performance remains inferior to that of Ex-Sliced-Calib.

Additionally, we examine the computational time of the BO-based calibration methods. In general, Ex-Sliced-Calib substantially reduces computation time compared to Full-Calib

TABLE 3

Comparison of test MSEs and calibrated parameter values from 20 experiments in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font)

Example	Method	Test MSE f	Calibrated parameter values		
			θ_1	θ_2	θ_3
I	Ex-Sliced-Calib	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	Rand-Sliced-Calib	13.935 (36.746)	420.308 (2.758)	348.473 (223.151)	—
	Full-Calib	92.890 (69.778)	278.484 (292.822)	311.888 (266.535)	—
	TR-Calib	71.931 (70.363)	169.509 (355.454)	337.334 (224.104)	—
	True parameter values		420.969	420.969	—
II	Ex-Sliced-Calib	0.355 (0.295)	0.018 (0.064)	− 0.013 (0.047)	—
	Rand-Sliced-Calib	0.562 (0.577)	0.006 (0.065)	0.024 (0.101)	—
	Full-Calib	1.807 (1.136)	0.018 (0.219)	0.060 (0.431)	—
	TR-Calib	0.783 (1.153)	−0.099 (0.293)	−0.041 (0.381)	—
	True parameter values		0	0	—
III	Ex-Sliced-Calib	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	Rand-Sliced-Calib	0.010 (0.031)	0.116 (0.016)	0.555 (0.008)	0.850 (0.009)
	Full-Calib	0.072 (0.050)	0.251 (0.191)	0.558 (0.023)	0.849 (0.016)
	TR-Calib	0.028 (0.033)	0.281 (0.151)	0.554 (0.002)	0.853 (0.001)
	True parameter values		0.115	0.556	0.853

and TR-Calib, although it remains more time-consuming than Rand-Sliced-Calib due to the additional surrogate modeling in the original parameter space to compute sensitivity indices. However, this cost may be negligible compared with the cost of a simulation run using the (moderately) expensive simulator employed in this study and in many applications. Also, the additional computational effort is offset by the significant efficiency gained through strategic dimensionality reduction. Further details can be found in the Supplementary Material (Jeong and Byon (2026)).

We also introduce a new metric, termed the normalized importance-weighted mean squared error (IMSE), which evaluates how well the proposed approach calibrates parameters while accounting for their relative importance, defined as

$$(18) \quad \text{IMSE} = \sum_{k \in \mathcal{K}^I} \omega_k \left(\frac{\theta_k - \hat{\theta}_k}{\text{range}(\theta_k)} \right)^2,$$

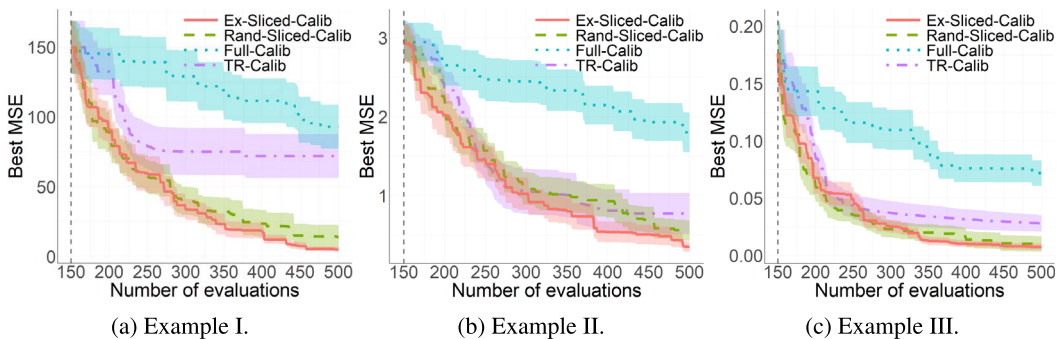


FIG. 5. Trajectories of the lowest MSEs with different methods in the numerical study (Note: the band denotes the ± 1 standard error).

TABLE 4

Comparison of IMSEs ($\times 100$) for the estimated parameters from 20 experiments in the numerical study (Note: The values inside parentheses are standard deviations, and the best results are highlighted in bold font)

Method	IMSE ($\times 100$)		
	Example I	Example II	Example III
Ex-Sliced-Calib	0.002 (0.003)	0.002 (0.002)	0.010 (0.012)
Rand-Sliced-Calib	2.657 (8.177)	0.003 (0.005)	0.012 (0.030)
Full-Calib	9.156 (12.814)	0.056 (0.089)	0.471 (0.715)
TR-Calib	11.821 (13.199)	0.054 (0.095)	0.369 (0.427)

where ω_k denotes the true importance measure obtained from the Sobol’ indices, θ_k the true value of the k th parameter, and $\hat{\theta}_k$ its estimated value, for $k \in \mathcal{K}^I$. Here $\text{range}(\theta_k)$ represents the positive difference between the upper and lower bounds of the allowable range of θ_k .

Table 4 compares the IMSE values of the proposed method with those of alternative approaches. The results demonstrate the clear superiority of Ex-Sliced-Calib, as this metric reflects not only the accuracy of parameter estimates but also their relative importance, independent of loss function values. Since only Ex-Sliced-Calib incorporates importance quantification of parameters during calibration, the IMSE metric provides a more informative measure of calibration accuracy in parameter estimation.

5.3. *Additional results.* We examine the influence of various algorithmic parameters, namely, d , p , and σ_0 , on the performance of Ex-Sliced-Calib. The Supplementary Material (Jeong and Byon (2026)) presents detailed sensitivity analysis results, which serve as the basis for setting guidelines for determining their values. Furthermore, to demonstrate how we quantify the uncertainty of the estimated parameters, we provide their confidence intervals and empirical coverage rates in the Supplementary Material (Jeong and Byon (2026)).

6. **Case study: Building energy simulation.** The performance of Ex-Sliced-Calib is assessed using hourly electricity consumption data from a residential building in the Mueller neighborhood area of Austin, Texas, in 2014. The building we analyze has a rectangular shape, with dimensions of $14 \times 9 \text{ m}^2$ on the first floor and $10 \times 9 \text{ m}^2$ on the second floor. This building reflects a common residential archetype in the *Pecan Street* dataset, which was collected through extensive data collection efforts using high-resolution sensors. More details about this dataset are provided in Section 2. The building’s size and configuration follow standard DOE reference settings, ensuring consistency and comparability with other studies in the energy modeling literature. We also select the year 2014 because its weather profile is included in the Typical Meteorological Year 3 dataset, enabling reproducible experiments.

To simulate the building’s electricity demand, we use EnergyPlus 9.3.0 (U.S. Department of Energy (2019)) as a BEM. We initialize the BEM with BEopt 2.8.0.0 (Christensen, Horowitz and U.S. Department of Energy Office of Energy Efficiency and Renewable Energy (2011)), which is an EnergyPlus-based software for evaluating building designs.

The meteorological station located near the building collects weather conditions. This information is then compiled into the EnergyPlus Weather File with a one-hour resolution. The compiled weather information is considered an observable input for EnergyPlus. The Input Data File, which contains adjustable calibration parameters, serves as another input data. The parameters are related to heating, ventilation, cooling, lighting, plug loads, etc. A total of 30 parameters are employed in this study, and some of them are listed in Table 1. Utilizing the two input files, we run simulations using EnergyPlus. A one-year simulation can typically be executed in around two minutes on a standard desktop computer.

TABLE 5

Comparison of training and test MSEs for different BO alternatives in the BEM case study (Note: The values in parentheses denote standard deviations, and the best results are shown in bold)

Method	Training MSE [kWh ²]	Test MSE [kWh ²]
Ex-Sliced-Calib	0.292 (0.002)	0.337 (0.004)
Rand-Sliced-Calib	0.298 (0.007)	0.344 (0.007)
Full-Calib	0.319 (0.010)	0.364 (0.014)
TR-Calib	0.312 (0.014)	0.354 (0.013)

6.1. *Comparison with other BO alternatives.* To assess the effectiveness of the proposed method in comparison to other alternatives, we randomly divide the hourly actual and simulated consumption data into a training set, which comprises approximately 70% of the entire weeks in 2014, and a test set consisting of the remaining weeks. The experiment is replicated 10 times for each method, where each set comprises 60 initial design points. The maximum number of EnergyPlus evaluations $N_{\max}$ is fixed at 300. For both Ex-Sliced-Calib and Rand-Sliced-Calib, we set the trial dimension as $d = 10$.

Table 5 summarizes the MSEs, and Figure 6 shows the MSE trajectories for each method with ± 1 standard error bands. Ex-Sliced-Calib consistently outperforms all alternatives across rounds. More specifically, as depicted in Figure 6(a), Ex-Sliced-Calib achieves a rapid decrease in MSE during its initial iterations, demonstrating the effectiveness of its strategy for prioritizing and selecting influential parameters more frequently. In contrast, Full-Calib, which operates in the original parameter space, exhibits a slow and gradual reduction in MSE across rounds. TR-Calib, while providing modestly better results than Full-Calib, still falls short compared to Ex-Sliced-Calib, as it also proceeds in the original parameter space.

Notably, Ex-Sliced-Calib holds a significant advantage over other approaches, as it can identify influential parameters and discern their importance. Figure 6(b) illustrates the trajectories of the posterior means in parameter weight distributions. The posterior mean of θ_4 's weight distribution surges above others, signifying its utmost influence. Additionally, θ_1 , θ_2 , and θ_8 show relatively larger posterior means of weight distributions than the remaining ones. In the BEM case study, these parameters correspond to *solar transmittance* (θ_1), *lighting level* (θ_2), *gross rated cooling COP* (θ_4), and *cooling supply air flow rate* (θ_8). This implies that this small subset of parameters exerts a dominant influence on energy consumption. In contrast, parameters, such as the *radiant fraction* (θ_{22}), exhibit only secondary or negligible effects in this study. These results not only validate the effect sparsity assumption but also enhance

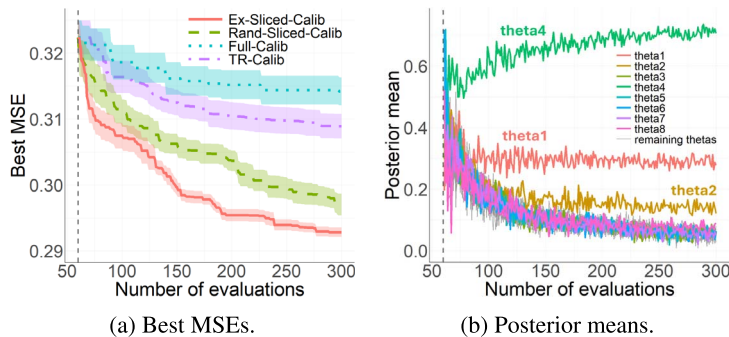


FIG. 6. (a) Trajectories of the lowest MSEs with different methods and (b) Trajectories of the posterior mean in the weight distributions of Ex-Sliced-Calib in the BEM case study.

model interpretability and enable a focus on the most important parameters, which is essential for practitioners working with BEMs involving a large number of parameters.

The identified important parameters can provide valuable insights when the BEM is used as a decision-making tool for building design and operations. In practice, energy auditors and engineers often face the challenge of calibrating a large number of parameters for a specific building's BEM with observational data. Ex-Sliced-Calib addresses this challenge by concentrating computational resources on the most influential parameters, resulting in more efficient calibration. This directly benefits tasks such as retrofit planning by prioritizing high-impact upgrades (e.g., replacing low-efficiency chillers or improving glazing with low solar transmittance), designing demand response strategies by adjusting cooling supply air flow rates during peak demand hours, and informing carbon reduction policies by identifying building parameter changes that maximize CO₂ reduction per dollar invested for compliance with local energy codes and climate action plans (Pallonetto et al. (2020), Dalla Valle et al. (2025)). Moreover, Ex-Sliced-Calib's demonstrated performance for EnergyPlus suggests its adaptability to other complex simulators such as TRNSYS (Klein et al. (2017)), Modelica (Modelica Association (2023)), and grid-level energy dispatch models.

Finally, recall model misspecification occurs when *cooling supply air flow rate*, one of the important parameters, is set close to its boundary value, as shown in Figure 1(a). Additionally, *fan total efficiency* (θ_{12}) turns out to be less important, yet also leads to model misspecification on its boundary values (see Figure 1(b)). This implies that model misspecification should be avoided regardless of parameter importance. Our slicing approach mitigates this issue by effectively balancing exploration and exploitation in the reduced-dimensional space.

6.2. Comparison with other calibration approaches. We further consider two well-regarded calibration methods: L_2 calibration (Tuo and Wu (2015)) and Bayesian calibration (Kennedy and O'Hagan (2001)). For L_2 calibration, given the considerable time it takes to process hourly data, we utilize data aggregated on daily and weekly bases. For Bayesian calibration, due to its intensive computational demands, we use only weekly aggregated data. We experiment with two prior specifications for parameters: uniform and Gaussian. For the Gaussian priors, we set the means to the default values in the BEM with a standard deviation of 0.2. To explore posteriors, the No-U-Turn Sampler is employed, a Hamiltonian Monte Carlo-based sampling technique effective in extracting high-dimensional posterior samples (Chong and Menberg (2018)). MCMC is configured to run 2,000 iterations with four chains. The first half of the samples serves as burn-in, while the second half is used for exploring the posteriors.

Table 6 summarizes the results. For comparison, we calculate metrics using the hourly data for all methods. Clearly, the MSEs from Ex-Sliced-Calib are significantly smaller than those of L_2 and Bayesian calibration. One key reason is the data resolution used for calibration. Another reason is that L_2 and Bayesian calibration depend solely on predesigned points to construct their surrogates, while Ex-Sliced-Calib adaptively chooses the next design point. Also, the role of the surrogate model differs among the techniques. The surrogate in Ex-Sliced-Calib directly approximates the MSE of the model output, whereas the surrogates in L_2 and Bayesian calibration are used to emulate the BEM responses or the physical process.

Even with the conceptual distinction between sequential and nonsequential design, its importance is particularly critical for the calibration problem in this work due to the high dimensionality and computational cost of the underlying simulation models. Non-sequential (one-shot) methods, such as L_2 or Bayesian calibration, must commit their simulation budget upfront, without feedback from intermediate results. As a consequence, a substantial fraction of simulations may be allocated to regions of the parameter space that are weakly informative or irrelevant for calibration accuracy, leading to poor scalability as the dimension increases.

TABLE 6

Comparison of the training and test MSEs for different calibration approaches in the BEM case study (Note: the values in parentheses denote standard deviations, and the best results are shown in bold)

Method (Prior)	Training data resolution	Training MSE [kWh ²]	Test MSE [kWh ²]
Ex-Sliced-Calib	hourly	0.292 (0.002)	0.337 (0.004)
L_2 calibration	weekly	1.482 (3.498)	1.650 (3.822)
L_2 calibration	daily	0.465 (0.162)	0.500 (0.148)
Bayesian calibration (uniform)	weekly	0.365 (0.012)	0.404 (0.017)
Bayesian calibration (Gaussian)	weekly	0.373 (0.015)	0.402 (0.012)

In contrast, Ex-Sliced-Calib employs sequential design as a necessity rather than a convenience: intermediate calibration results are actively used to identify influential parameters and progressively reduce the effective dimension of the problem. By slicing the original parameter space according to estimated parameter importance, it focuses new simulation queries on lower-dimensional, informative subspaces. This adaptive allocation of computational effort is essential for making the calibration problem tractable under realistic simulation budgets. Moreover, designing query points within these importance-driven slices mitigates practical issues, such as excessive sampling near the boundaries of the parameter space, which can exacerbate model misspecification and degrade calibration performance (see Figure 1).

Lastly, we note that the practical importance of computational efficiency arises not from a single calibration instance but from the scale and decision context in which BEM calibration is deployed. Although the numerical experiments in this study focus on a single building due to limited data availability, the motivating application is large-scale BEM deployment across many buildings. For example, the *Pecan Street Project* alone includes approximately 300 buildings (Rowan (2022)), and each building requires its own calibrated model due to heterogeneous building types and uncertainty scenarios. Even conservative savings of a few hours per building translate into several weeks of wall-clock time saved across a community-scale deployment. At the scale of the broader Mueller development, planned to serve roughly 6,900 residents (Mueller Foundation (2023)), this reduction accumulates to many months of computational time. At this scale, computational efficiency becomes a binding constraint on whether large-scale calibration is feasible at all, rather than merely convenient.

Moreover, shorter calibration cycles support more efficient downstream decision-making, such as retrofitting, planning, or deploying demand response programs. When calibration is computationally burdensome, it restricts the ability to explore alternative policies or conduct meaningful what-if analyses. If calibration costs dominate the decision process, the benefits of calibration could diminish, potentially yielding limited or even negligible practical gains.

7. Conclusion. This study devises a new importance-driven calibration methodology designed for cases where a subset of parameters are influential and their levels of importance differ. The proposed method reduces the dimensionality with parameter selection for effective exploration by leveraging imbalanced effects among parameters. This selection is guided by the weight distributions, which also explain the parameter importance. These weight distributions, therefore, serve a dual purpose: they direct the calibration focus toward parameters of greater significance while also providing a quantitative measure of parameter importance.

In the future we will extend our methodology to incomplete computer models with model bias (Jeong and Byon (2024)) and to functional calibration (Ezzat, Pourhabib and Ding (2018)). We will also develop collaborative and federated learning approaches for calibrating parameters across multiple systems (Kontar et al. (2021), Jeong, Yue and Chung (2026)).

8. Significance statement. Physics-based computer models often require the specification of key parameters, and estimating these values from real-world data is critical in many scientific and engineering domains. As the number of parameters grows, existing calibration methods often become unreliable and computationally expensive. This study proposes an importance-driven calibration approach that identifies and prioritizes parameters with the greatest influence on simulation outputs. By concentrating computational effort on influential parameters, the proposed method enhances calibration efficiency, robustness, and interpretability, making it more scalable and accessible. A case study on building energy simulation demonstrates the potential of the approach to support operational decision-making, such as prioritizing retrofitting investments and designing effective demand response programs.

Funding. This work was supported in part by the U.S. National Science Foundation under Grant No. CMMI-2226348.

SUPPLEMENTARY MATERIAL

Supplement (DOI: [10.1214/26-AOAS2148SUPPA](https://doi.org/10.1214/26-AOAS2148SUPPA); .pdf). Proofs and additional numerical results (Jeong and Byon (2026)).

Code (DOI: [10.1214/26-AOAS2148SUPPB](https://doi.org/10.1214/26-AOAS2148SUPPB); .zip). R codes for numerical examples are available at <https://github.com/UM-RAISE>.

REFERENCES

- BA, S., MYERS, W. R. and WANG, D. (2018). A sequential maximum projection design framework for computer experiments with inert factors. *Statist. Sinica* **28** 879–897. [MR3791092](#)
- BADANIDIYURU, A., KLEINBERG, R. and SLIVKINS, A. (2018). Bandits with knapsacks. *J. ACM* **65** 1–55. [MR3771540](#) <https://doi.org/10.1145/3164539>
- BINOIS, M. and WYCOFF, N. (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Trans. Evol. Learn. Optim.* **2**.
- BOOTH, A., CHOUDHARY, R. and SPIEGELHALTER, D. (2013). A hierarchical Bayesian framework for calibrating micro-level models with macro-level data. *J. Build. Perform. Simul.* **6** 293–318.
- CALAMA-GONZÁLEZ, C. M., SYMONDS, P., PETROU, G., SUÁREZ, R. and LEÓN-RODRÍGUEZ, Á. L. (2021). Bayesian calibration of building energy models for uncertainty analysis through test cells monitoring. *Appl. Energy* **282** 116118.
- CESA-BIANCHI, N. and LUGOSI, G. (2012). Combinatorial bandits. *J. Comput. System Sci.* **78** 1404–1422. [MR2926141](#) <https://doi.org/10.1016/j.jcss.2012.01.001>
- CHAKRABARTY, A., MADDALENA, E., QIAO, H. and LAUGHMAN, C. (2021). Scalable Bayesian optimization for model calibration: Case study on coupled building and HVAC dynamics. *Energy Build.* **253** 111460.
- CHEN, X., LI, K., MI, Z. and WANG, F. (2024). An air conditioning control method based on non-measurement thermal comfort standard correction for demand response. *IEEE Trans. Ind. Appl.* **60** 4736–4748.
- CHONG, A. and MENBERG, K. (2018). Guidelines for the Bayesian calibration of building energy models. *Energy Build.* **174** 527–547.
- CHRISTENSEN, C., HOROWITZ, S. and U.S. DEPARTMENT OF ENERGY OFFICE OF ENERGY EFFICIENCY AND RENEWABLE ENERGY (2011). BEopt (Building Energy Optimization Tool) [SWR-05-41], Vers. 2.8.0.0.
- COAKLEY, D., RAFTERY, P. and KEANE, M. (2014). A review of methods to match building energy simulation models to measured data. *Renew. Sustain. Energy Rev.* **37** 123–141.
- DALLA VALLE, A., TOOSI, H. A., LEONFORTE, F., DEL PERO, C., LAVAGNA, M., CAMPIOLI, A. and ASTE, N. (2025). Measuring the impact of holistic energy retrofit strategies: Life cycle assessment aligned with level (s). *Energy Build.* 116357.
- DINH, H. T., LEE, K.-H. and KIM, D. (2022). Supervised-learning-based hour-ahead demand response for a behavior-based home energy management system approximating MILP optimization. *Appl. Energy* **321** 119382.
- DIOUANE, Y., PICHENY, V., LE RICHE, R. and SCOTTO DI PERROTOLO, A. (2023). TREGO: A trust-region framework for efficient global optimization. *J. Global Optim.* **86** 1–23. [MR4578206](#) <https://doi.org/10.1007/s10898-022-01245-w>

- ERIKSSON, D., PEARCE, M., GARDNER, J. R., TURNER, R. and POLOCZEK, M. (2019). Scalable global optimization via local Bayesian optimization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.
- EZZAT, A. A., POURHABIB, A. and DING, Y. (2018). Sequential design for functional calibration of computer models. *Technometrics* **60** 286–296. [MR3847166](#) <https://doi.org/10.1080/00401706.2017.1377638>
- FENG, Y., DUAN, Q., CHEN, X., YAKKALI, S. S. and WANG, J. (2021). Space cooling energy usage prediction based on utility data for residential buildings using machine learning methods. *Appl. Energy* **291** 116814.
- GONZÁLEZ-TORRES, M., PÉREZ-LOMBARD, L., CORONEL, J. F., MAESTRE, I. R. and YAN, D. (2022). A review on buildings energy information: Trends, end-uses, fuels and drivers. *Energy Rep.* **8** 626–637.
- GRAMACY, R. B. (2020). *Surrogates—Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR4283556](#) <https://doi.org/10.1201/9780367815493>
- HIGDON, D., KENNEDY, M., CAVENDISH, J. C., CAFFEO, J. A. and RYNE, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.* **26** 448–466. [MR2116355](#) <https://doi.org/10.1137/S1064827503426693>
- HOMMA, T. and SALTELLI, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* **52** 1–17.
- IOOSS, B. and LEMAÎTRE, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications* 101–122.
- JAIN, P., SHASHAANI, S. and BYON, E. (2023). Wake effect parameter calibration with large-scale field operational data using stochastic optimization. *Appl. Energy* **347** 121426.
- JAIN, P., SHASHAANI, S. and BYON, E. (2025). Simulation model calibration with dynamic stratification and adaptive sampling. *J. Simul.* **19** 494–515.
- JEONG, C. and BYON, E. (2024). Calibration of building energy computer models via bias-corrected iteratively reweighted least squares method. *Appl. Energy* **360** 122753.
- JEONG, C. and BYON, E. (2026). Supplement to “Explainable parameter calibration via importance-driven sequential design with an application to building energy systems.” <https://doi.org/10.1214/26-AOAS2148SUPPA>, <https://doi.org/10.1214/26-AOAS2148SUPPB>
- JEONG, C., YUE, X. and CHUNG, S. (2026). Fed-joint: Joint modeling of nonlinear degradation signals and failure events for remaining useful life prediction using federated learning. *Reliab. Eng. Syst. Saf.* **267** 111833.
- JEONG, C., ZIANG, X., BYON, E., BERAHAS, A. S. and CETIN, K. (2023). Multiblock parameter calibration in computer models. *INFORMS J. Data Sci.* **2** 116–137.
- JONES, D. R., SCHONLAU, M. and WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *J. Global Optim.* **13** 455–492. [MR1673460](#) <https://doi.org/10.1023/A:1008306431147>
- JOSEPH, V. R., GUL, E. and BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102** 371–380. [MR3371010](#) <https://doi.org/10.1093/biomet/asv002>
- KENNEDY, M. C. and O’HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. [MR1858398](#) <https://doi.org/10.1111/1467-9868.00294>
- KLEIN, S. A. et al. (2017). *TRNSYS 18: A Transient System Simulation Program Solar Energy Laboratory*. Univ. Wisconsin Press, Madison, USA.
- KONTAR, R., SHI, N., YUE, X., CHUNG, S., BYON, E., CHOWDHURY, M., JIN, J., KONTAR, W., MASOUD, N. et al. (2021). The Internet of federated things (IoFT). *IEEE Access* **9** 156071–156113.
- LI, C., GUPTA, S., RANA, S., NGUYEN, V., VENKATESH, S. and SHILTON, A. (2017). High-dimensional Bayesian optimization using dropout. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* 2096–2102.
- LI, Q., AUGENBROE, G. and BROWN, J. (2016). Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. *Energy Build.* **124** 194–202.
- LI, S., KO, Y. M. and BYON, E. (2021). Nonparametric importance sampling for wind turbine reliability analysis with stochastic computer models. *Ann. Appl. Stat.* **15** 1850–1871. [MR4355079](#) <https://doi.org/10.1214/21-aos1490>
- LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and YE, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490. [MR2328617](#) <https://doi.org/10.1198/004017006000000228>
- LIU, B., YUE, X., BYON, E. and KONTAR, R. A. (2022). Parameter calibration in wake effect simulation model with stochastic gradient descent and stratified sampling. *Ann. Appl. Stat.* **16** 1795–1821. [MR4455900](#) <https://doi.org/10.1214/21-aos1567>
- LIU, Y., DONG, Z., LIU, B., XU, Y. and DING, Z. (2023). FedForecast: A federated learning framework for short-term probabilistic individual load forecasting in smart grid. *Int. J. Electr. Power Energy Syst.* **152** 109172.

- MA, J., ZHANG, J., LI, R., ZHENG, H. and LI, W. (2022). Using Bayesian optimization to automate the calibration of complex hydrological models: Framework and application. *Environ. Model. Softw.* **147** 105235.
- MARTINEZ-VIOL, V., URBANO, E. M., DELGADO-PRIETO, M. and ROMERAL, L. (2022). Automatic model calibration for coupled HVAC and building dynamics using Modelica and Bayesian optimization. *Build. Environ.* **226** 109693.
- MOČKUS, J. (1975). On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference* 400–404. Springer, Berlin.
- MODELICA ASSOCIATION (2023). Modelica—a Unified Object-Oriented Language for Systems Modeling, Language Specification Version 3.6. <https://specification.modelica.org/maint/3.6/MLS.pdf>.
- MUELLER FOUNDATION (2023). Mueller Affordable Homes Program: Frequently Asked Questions. <https://muelleraustin.com/wp-content/uploads/2023/08/Mueller-Affordable-Homes-FAQ-1.pdf>.
- PALLONETTO, F., DE ROSA, M., D'ETTORRE, F. and FINN, D. P. (2020). On the assessment and control optimisation of demand response programs in residential buildings. *Renew. Sustain. Energy Rev.* **127** 109861.
- PARK, J., BYON, E., KO, Y. M. and SHASHAANI, S. (2025). Strata design for variance reduction in stochastic simulation. *Technometrics* **67** 203–214. MR4902935 <https://doi.org/10.1080/00401706.2024.2416411>
- PECAN STREET INC. (2025). Pecan Street Inc. <https://www.pecanstreet.org/>. Accessed: 2025-08-15.
- RAMOS, P. V. B., VILLELA, S. M., SILVA, W. N. and DIAS, B. H. (2023). Residential energy consumption forecasting using deep learning models. *Appl. Energy* **350** 121705.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435
- REGIS, R. G. (2016). Trust regions in Kriging-based optimization with expected improvement. *Eng. Optim.* **48** 1037–1059. MR3473811 <https://doi.org/10.1080/0305215X.2015.1082350>
- REIKER, T., GOLUMBEANU, M., SHATTOCK, A., BURGERT, L., SMITH, T. A., FILIPPI, S., CAMERON, E. and PENNY, M. A. (2021). Emulator-based Bayesian optimization for efficient multi-objective calibration of an individual-based model of malaria. *Nat. Commun.* **12** 7212.
- ROWAN, C. (2022). Pecan Street on KXAN: Why One Austin Neighborhood Tracks Energy Use down to the Circuit. <https://www.pecanstreet.org/2022/07/kxan/>.
- RUSSO, D., VAN ROY, B., KAZEROONI, A., OSBAND, I., WEN, Z. et al. (2018). A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **11** 1–96.
- RUTTER, C. M., OZIK, J., DEYOREO, M. and COLLIER, N. (2019). Microsimulation model calibration using incremental mixture approximate Bayesian computation. *Ann. Appl. Stat.* **13** 2189–2212. MR4037427 <https://doi.org/10.1214/19-aos1279>
- SALEM, M. B., BACHOC, F., ROUSTANT, O., GAMBOA, F. and TOMASO, L. (2019). Gaussian process-based dimension reduction for goal-oriented sequential design. *SIAM/ASA J. Uncertain. Quantificat.* **7** 1369–1397. MR4041716 <https://doi.org/10.1137/18M1167930>
- SALTELLI, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.* **145** 280–297.
- SALTELLI, A., TARANTOLA, S. and CHAN, K. P. S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* **41** 39–56.
- SHA, D., OZBAY, K. and DING, Y. (2020). Applying Bayesian optimization for calibration of transportation simulation models. *Transp. Res. Rec.* **2674** 215–228.
- SHEN, Y. and KINGSFORD, C. (2023). Computationally efficient high-dimensional Bayesian optimization via variable selection. In *Proceedings of the Second International Conference on Automated Machine Learning. Proceedings of Machine Learning Research* **224** 15/1–27.
- SOBOL, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **55** 271–280. MR1823119 [https://doi.org/10.1016/S0378-4754\(00\)00270-6](https://doi.org/10.1016/S0378-4754(00)00270-6)
- SOKOL, J., DAVILA, C. C. and REINHART, C. F. (2017). Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy Build.* **134** 11–24.
- SPAGNOL, A., LE RICHE, R. and DA VEIGA, S. (2019). Global sensitivity analysis for optimization with variable selection. *SIAM/ASA J. Uncertain. Quantificat.* **7** 417–443. MR3939340 <https://doi.org/10.1137/18M1167978>
- SRINIVAS, N., KRAUSE, A., KAKADE, S. M. and SEEGER, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning* 1015–1022.
- SURJANOVIC, S. and BINGHAM, D. (2013). Virtual library of simulation experiments: Test functions and datasets. <https://www.sfu.ca/~ssurjano/>.
- TUO, R. and WU, C. F. J. (2015). Efficient calibration for imperfect computer models. *Ann. Statist.* **43** 2331–2352. MR3405596 <https://doi.org/10.1214/15-AOS1314>
- U. S. DEPARTMENT OF ENERGY (2019). EneryPlus essentials Technical Report.
- U. S. ENERGY INFORMATION ADMINISTRATION (2015). Table 2.1 Energy Consumption by Sectors.

- WINKEL, M. A., STALLRICH, J. W., STORLIE, C. B. and REICH, B. J. (2021). Sequential optimization in locally important dimensions. *Technometrics* **63** 236–248. [MR4251497](#) <https://doi.org/10.1080/00401706.2020.1714738>
- WYCOFF, N., BINOIS, M. and WILD, S. M. (2021). Sequential learning of active subspaces. *J. Comput. Graph. Statist.* **30** 1224–1237. [MR4356616](#) <https://doi.org/10.1080/10618600.2021.1874962>
- XU, Z., JEONG, C., BYON, E. and CETIN, K. (2021). Season-dependent parameter calibration in building energy simulation. In *Proceedings of 2021 IISE Annual Conference* 423–428.
- YANG, Y., JI, C. and DENG, K. (2021). Rapid design of metamaterials via multitarget Bayesian optimization. *Ann. Appl. Stat.* **15** 768–796. [MR4298963](#) <https://doi.org/10.1214/20-aos1426>
- YUAN, J. and NG, S. H. (2020). An integrated method for simultaneous calibration and parameter selection in computer models. *ACM Trans. Model. Comput. Simul.* **30** 1–23. [MR4066848](#) <https://doi.org/10.1145/3364217>
- ZHAN, S., WICHERN, G., LAUGHMAN, C., CHONG, A. and CHAKRABARTY, A. (2022). Calibrating building simulation models using multi-source datasets and meta-learned Bayesian optimization. *Energy Build.* **270** 112278.

SUPPLEMENTARY MATERIAL: EXPLAINABLE PARAMETER CALIBRATION VIA IMPORTANCE-DRIVEN SEQUENTIAL DESIGN WITH AN APPLICATION TO BUILDING ENERGY SYSTEMS

BY CHEOLJOON JEONG AND EUNSHIN BYON

A. Proof of Theorem 4.1. This section provides the proof of Theorem 4.1. To facilitate the regret bound analysis, we introduce a simplified version of the Ex-Sliced-Calib algorithm, as summarized in Algorithm 1. Unlike the original algorithm that dynamically learns parameter importance, Algorithm 1 assumes the knowledge of influential parameters *a priori*. This simplification makes it tractable to derive a regret bound and provides theoretical insight into the impact of effect sparsity and imbalance on the performance of the BO procedure.

Algorithm 1 Simplified Ex-Sliced-Calib

- 1: **Input:** $N_0 = 0$, (i.e., $\mathcal{D}_{N_0} = \emptyset$), $N_{\max} (> N_0)$, θ_0^{NI} , d_I , $\mathcal{K}^I$, and $\mathcal{K}^{NI}$
 - 2: **for** $n = 1, \dots, N$ **do**
 - 3: Build the GP surrogate $\hat{f}^I(\theta^I)$ with samples $\mathcal{D}_{n-1}^I$.
 - 4: Design the major parameters θ_n^I by utilizing

$$\theta_n^I = \arg \min_{\theta^I \in \Theta^I} \mu((\theta^I, \theta_0^{NI}) | \mathcal{D}_{n-1}) - \sqrt{\beta_n} \sigma((\theta^I, \theta_0^{NI}) | \mathcal{D}_{n-1}),$$
 where $\mu((\theta^I, \theta_0^{NI}) | \mathcal{D}_{n-1})$ and $\sigma((\theta^I, \theta_0^{NI}) | \mathcal{D}_{n-1})$ are the mean and standard deviation of the GP at θ^I , respectively.
 - 5: Design the minor parameters $\theta_n^{NI} = \theta_0^{NI}$.
 - 6: Set the new query design θ_n by combining θ_n^I and θ_n^{NI} , i.e., $\theta_n = (\theta_n^I, \theta_n^{NI})$.
 - 7: Evaluate f at θ_n and set $\mathcal{D}_n = \mathcal{D}_{n-1} \cup (\theta_n, f(\theta_n))$.
 - 8: **end for**
 - 9: **Output:** $(\hat{\theta}, f(\hat{\theta})) = (\theta_{\min}, f(\theta_{\min}))$.
-

We begin by defining the axis-aligned sparse loss function $f(\cdot)$ and make some assumptions.

DEFINITION A.1 (Axis-aligned sparse function). *Let $\Theta := [0, 1]^D$. A function $f : \Theta \rightarrow \mathbb{R}$ is an axis-aligned sparse function if there exists $\theta^I \in \Theta^I = [0, 1]^{d_I}$ and $\theta^{NI} \in \Theta^{NI} = [0, 1]^{D-d_I}$ with $0 < d_I < D$ such that $f(\theta) \simeq f(\theta^I, \theta^{NI})$ for any θ^{NI} (here, $\theta := (\theta^I, \theta^{NI})$ is expressed without an order of elements and “ $\simeq$ ” indicates that θ^{NI} have little affect f).*

ASSUMPTION A.1. *Suppose that (A1) the number of entire dimensions D is finite and known; (A2) the computer model is deterministic so the same outputs are obtained with the same inputs; (A3) f is calculated without noise; (A4) f is an axis-aligned sparse function; (A5) f is bounded over Θ , i.e., there exists $B \in \mathbb{R}_+$ such that $|f(\theta)| < B$ for all $\theta \in \Theta$; (A6) Θ is compact and convex; (A7) at least one minimizer θ^* lies in a continuous region with positive measure; (A8) a kernel function is smooth.*

In Assumption A.1, A1 typically holds because we have the information about calibration parameters through the description of a simulator. A2 is satisfied with the deterministic

computer model such as the BEM considered in our case study. Also, field observations are usually given so we consider that the loss function f is not subject to noise, satisfying A3. A4 holds for the problem structure considered in this study. A5 to A7 are generally satisfied. A8 holds with squared exponential and Mart rn kernels that are popularly used in BO.

To reflect the effect sparsity and effect imbalance, we establish the following assumption that the absolute values of partial derivatives of $f(\cdot)$ on each $D - d_I$ parameter in $\mathcal{K}^{NI}$ are much smaller than those on parameters in $\mathcal{K}^I$ and that their degrees of influence are different.

ASSUMPTION A.2. *Suppose that Assumption A.1 holds. Let $f(\boldsymbol{\theta})$ be a sample path drawn from a GP with mean zero, which satisfies the following probability bounds on each partial derivative of $f(\boldsymbol{\theta})$ with respect to θ_k for $k = 1, \dots, D$ with some constants $b, c > 0$ as follows.*

$$(A.1) \quad P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| > L\right) \leq c \cdot \exp\left[-\left(\frac{L}{\alpha_k b}\right)^2\right], \quad \forall k \in \mathcal{K}^I,$$

$$(A.2) \quad P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| > L\right) \leq c \cdot \exp\left[-\left(\frac{L}{\rho_k b}\right)^2\right], \quad \forall k \in \mathcal{K}^{NI},$$

for some $L > 0$, where $0 \leq \{\rho_k : \forall k \in \mathcal{K}^{NI}\} \ll \{\alpha_k : \forall k \in \mathcal{K}^I\} \leq 1$.

The inequalities in (A.1) and (A.2) represent the key simplifications that distinguish the simplified Ex-Sliced-Calib from its original procedure. Here, α_k and ρ_k exhibit the extent to which θ_k affects f among the influential and non-influential parameters, respectively. As $\alpha_k \rightarrow 1$, θ_k becomes more influential, whereas θ_k becomes less influential as $\rho_k \rightarrow 0$. Using Assumptions A.1–A.2, we derive the regret bound.

First, Lemma A.1 below analyzes the bound of the instantaneous regret $r_n = f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*)$. Hereafter, the parameter domain $\boldsymbol{\Theta}$ is considered as $\boldsymbol{\Theta} := [0, 1]^D$. The result in Lemma A.1 implies that f is Lipschitz continuous with $L = b\sqrt{\log(2Dc/\delta)}$.

LEMMA A.1. *Suppose that Assumption A.2 holds. For all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$,*

$$(A.3) \quad |f(\boldsymbol{\theta}) - f(\tilde{\boldsymbol{\theta}})| \leq \alpha_{\max} L \|\boldsymbol{\theta}^I - \tilde{\boldsymbol{\theta}}^I\|_1 + \rho_{\max} L \|\boldsymbol{\theta}^{NI} - \tilde{\boldsymbol{\theta}}^{NI}\|_1$$

holds with probability $\geq 1 - \delta/2$, where $\delta = 2Dc \cdot \exp(-(L/b)^2)$ for the constants b and c that satisfy (A.1) and (A.2), with $L = b\sqrt{\log(2Dc/\delta)}$, $\alpha_{\max} = \max\{\alpha_k : \forall k \in \mathcal{K}^I\}$, and $\rho_{\max} = \max\{\rho_k : \forall k \in \mathcal{K}^{NI}\}$.

Proof. From Assumption A.2, we can rewrite (A.1) and (A.2) by substituting L with $\alpha_k L$ and $\rho_k L$, respectively, as follows.

$$(A.4) \quad P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| > \alpha_k L\right) \leq c \cdot \exp\left[-\left(\frac{L}{b}\right)^2\right], \quad \forall k \in \mathcal{K}^I,$$

$$(A.5) \quad P\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| > \rho_k L\right) \leq c \cdot \exp\left[-\left(\frac{L}{b}\right)^2\right], \quad \forall k \in \mathcal{K}^{NI},$$

for some constants $b, c > 0$ and $0 \leq \{\rho_k : \forall k \in \mathcal{K}^{NI}\} \ll \{\alpha_k : \forall k \in \mathcal{K}^I\} \leq 1$. Applying the union bound to all D events in (A.4) and (A.5), we have

$$(A.6) \quad P\left(\bigcup_{\forall k \in \mathcal{K}^I} \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| > \alpha_k L\right) \bigcup_{\forall k \in \mathcal{K}^{NI}} \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| > \rho_k L\right)\right) \leq Dc \cdot \exp\left[-\left(\frac{L}{b}\right)^2\right],$$

or equivalently,

$$(A.7) \quad P \left(\bigcap_{\forall k \in \mathcal{K}^I} \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| \leq \alpha_k L \right) \bigcap_{\forall k \in \mathcal{K}^{NI}} \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| \leq \rho_k L \right) \right) \geq 1 - Dc \cdot \exp \left[- \left(\frac{L}{b} \right)^2 \right].$$

Therefore, we have

$$(A.8) \quad \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| \leq \alpha_k L, \forall k \in \mathcal{K}^I \quad \text{and} \quad \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| \leq \rho_k L, \forall k \in \mathcal{K}^{NI},$$

which lead to

$$(A.9) \quad \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| \leq \alpha_k L, \forall k \in \mathcal{K}^I \quad \text{and} \quad \left| \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} \right| \leq \rho_k L, \forall k \in \mathcal{K}^{NI},$$

for all $\boldsymbol{\theta}$ with probability $\geq 1 - Dc \cdot \exp(-(L/b)^2)$.

On the other hand, by the mean value theorem, the following equalities hold with $\boldsymbol{\theta}_\varepsilon = (1 - \varepsilon)\boldsymbol{\theta} + \varepsilon\tilde{\boldsymbol{\theta}}$ for some $\varepsilon \in [0, 1]$.

$$(A.10) \quad \begin{aligned} f(\boldsymbol{\theta}) - f(\tilde{\boldsymbol{\theta}}) &= \nabla f(\boldsymbol{\theta}_\varepsilon)^T (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\ &= \sum_{k=1}^D \left[\frac{\partial f(\boldsymbol{\theta}_\varepsilon)}{\partial \theta_k} \right] (\theta_k - \tilde{\theta}_k), \quad \forall \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}. \end{aligned}$$

Taking absolute values on both sides in (A.10), we have

$$(A.11) \quad \begin{aligned} |f(\boldsymbol{\theta}) - f(\tilde{\boldsymbol{\theta}})| &= \left| \sum_{k=1}^D \left[\frac{\partial f(\boldsymbol{\theta}_\varepsilon)}{\partial \theta_k} \right] (\theta_k - \tilde{\theta}_k) \right| \\ &\leq \sum_{k=1}^D \left| \frac{\partial f(\boldsymbol{\theta}_\varepsilon)}{\partial \theta_k} \right| |\theta_k - \tilde{\theta}_k| \\ &\leq \sum_{\forall k \in \mathcal{K}^I} \alpha_k L |\theta_k - \tilde{\theta}_k| + \sum_{\forall k \in \mathcal{K}^{NI}} \rho_k L |\theta_k - \tilde{\theta}_k| \\ &\leq \alpha_{\max} L \|\boldsymbol{\theta}^I - \tilde{\boldsymbol{\theta}}^I\|_1 + \rho_{\max} L \|\boldsymbol{\theta}^{NI} - \tilde{\boldsymbol{\theta}}^{NI}\|_1, \quad \forall \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}, \end{aligned}$$

with probability $\geq 1 - Dc \cdot \exp(-(L/b)^2)$. The third inequality holds due to the fact that the absolute value of partial derivative of f is bounded above for all $\boldsymbol{\theta}$ due to (A.9). $\square$

Consider the discretized subsets of parameter domain, $\boldsymbol{\Theta}_n \subset \boldsymbol{\Theta}$, of size $|\boldsymbol{\Theta}_n| < \infty$, with $\alpha_{\max} \xi_n$ evenly spaced points for some ξ_n in the k th dimension for all $k \in \mathcal{K}^I$ and $\rho_{\max} \xi_n$ evenly spaced points in the k th dimension for all $k \in \mathcal{K}^{NI}$. Let $\boldsymbol{\theta}_{[n]}$ denote the closest point to $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}_n$ in terms of the L_1 distance. From Lemma A.1, we have

$$(A.12) \quad \begin{aligned} |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_{[n]})| &\leq \alpha_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \|\boldsymbol{\theta}^I - \boldsymbol{\theta}_{[n]}^I\|_1 + \rho_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \|\boldsymbol{\theta}^{NI} - \boldsymbol{\theta}_{[n]}^{NI}\|_1 \\ &\leq \alpha_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \frac{d_I}{\alpha_{\max} \xi_n} + \rho_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \frac{D - d_I}{\rho_{\max} \xi_n} \\ &= b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \frac{D}{\xi_n}, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \end{aligned}$$

with probability $\geq 1 - \delta/2$ for any positive integer n . The second inequality holds because the distance between θ_k and $\theta_{[n],k}$ is less than $1/(\alpha_{\max}\xi_n)$ for all $k \in \mathcal{K}^I$ and $1/(\rho_{\max}\xi_n)$ for all $k \in \mathcal{K}^{NI}$.

If we set $\xi_n = Dn^2b\sqrt{\log(2Dc/\delta)}$ in (A.12), we obtain

$$(A.13) \quad |f(\theta) - f(\theta_{[n]})| \leq \frac{1}{n^2}, \quad \forall \theta \in \Theta, \forall n \geq 1,$$

with probability $\geq 1 - \delta/2$.

Additionally, the size of Θ_n denoted by $|\Theta_n|$ is as follows.

$$(A.14) \quad \begin{aligned} |\Theta_n| &= (\alpha_{\max}\xi_n)^{d_I} (\rho_{\max}\xi_n + 1)^{D-d_I} \\ &= \left[\alpha_{\max} Dn^2b \sqrt{\log\left(\frac{2Dc}{\delta}\right)} \right]^{d_I} \left[\rho_{\max} Dn^2b \sqrt{\log\left(\frac{2Dc}{\delta}\right)} + 1 \right]^{D-d_I}, \end{aligned}$$

with probability $\geq 1 - \delta/2$. The reason we have “+1” in the second element in (A.14) is that we also have the fixed point θ_0^{NI} for the minor parameter design.

Meanwhile, let us consider the GP-LCB algorithm with a noise-free setting in Algorithm 2 to connect the simplified Ex-Sliced-Calib with the GP-LCB framework, where LCB and UCB stand for lower and upper confidence bounds, respectively. GP-LCB is the minimization version of GP-UCB, which is BO with the UCB acquisition function (Srinivas et al., 2010). Here, GP-LCB is used without any noise for $f(\cdot)$ because we aim to minimize the empirical loss $f(\cdot)$.

Algorithm 2 GP-LCB

- 1: **Input:** $N_0 = 0$, $N_{\max}$, and LCB acquisition function.
 - 2: $\hat{\mathcal{D}}_{N_0} = \emptyset$.
 - 3: **for** $n = 1, \dots, N$ **do**
 - 4: Build the GP surrogate $\hat{f}(\cdot)$ with samples $\hat{\mathcal{D}}_{n-1}$.
 - 5: Design the parameters by $\hat{\theta}_n = \arg \min_{\theta \in \Theta} \hat{\mu}(\theta|\hat{\mathcal{D}}_{n-1}) - \sqrt{\beta_n} \hat{\sigma}(\theta|\hat{\mathcal{D}}_{n-1})$.
 - 6: Evaluate f at $\hat{\theta}_n$ and set $\hat{\mathcal{D}}_n = \hat{\mathcal{D}}_{n-1} \cup (\hat{\theta}_n, f(\hat{\theta}_n))$
 - 7: **end for**
 - 8: **Output:** $(\hat{\theta}, f(\hat{\theta})) = (\hat{\theta}_{\min}, f(\hat{\theta}_{\min}))$.
-

Note that the simplified Ex-Sliced-Calib (Algorithm 1) and GP-LCB (Algorithm 2) design different query points at each round by constructing the different GPs with different predictive distributions. Algorithm 1 applies BO to the major parameters only, whereas Algorithm 2 operates on the entire parameter space. In the subsequent proof, we will use *hat* notations for the elements obtained from GP-LCB. To be concise, we will exclude the dataset symbol $\mathcal{D}_n$ and $\hat{\mathcal{D}}_n$ from $\mu(\cdot|\mathcal{D}_n)$, $\hat{\mu}(\cdot|\mathcal{D}_n)$, $\sigma(\cdot|\mathcal{D}_n)$, and $\hat{\sigma}(\cdot|\mathcal{D}_n)$, by representing $\mu_n(\cdot)$, $\hat{\mu}_n(\cdot)$, $\sigma_n(\cdot)$, and $\hat{\sigma}_n(\cdot)$, respectively, if no confusion arises.

Now, we present some useful lemmas to proceed further with deriving the regret bound.

LEMMA A.2 (Lemma 5.6 in Srinivas et al. (2010)). *Consider the subsets of parameter domain, $\Theta_n \subset \Theta$, of size $|\Theta_n| < \infty$. Pick $\delta \in (0, 1)$ and set $\beta_n = 2 \log(|\Theta_n|\omega_n/\delta)$ in Algorithm 2, where $\sum_{n \geq 1} \omega_n^{-1} = 1$ and $\omega_n > 0$. Then we have*

$$(A.15) \quad |f(\theta) - \hat{\mu}_{n-1}(\theta)| \leq \sqrt{\beta_n} \hat{\sigma}_{n-1}(\theta), \quad \forall \theta \in \Theta_n, \forall n \geq 1,$$

with probability $\geq 1 - \delta$.

The result in Lemma A.2 can be applied to the simplified Ex-Sliced-Calib to obtain

$$(A.16) \quad |f(\boldsymbol{\theta}) - \mu_{n-1}(\boldsymbol{\theta})| \leq \sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_n, \quad \forall n \geq 1,$$

with probability $\geq 1 - \delta$.

Using the size of the discretized $\boldsymbol{\Theta}_n$ in (A.14), β_n in (A.16) can be represented by

$$(A.17) \quad \begin{aligned} \beta_n &= 2 \log \left(\frac{|\boldsymbol{\Theta}_n| \omega_n}{\delta} \right) \\ &= 2 \log \left(\frac{\omega_n}{\delta} \right) + 2d_I \log \left[\alpha_{\max} D n^2 b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \right] \\ &\quad + 2(D - d_I) \log \left[\rho_{\max} D n^2 b \sqrt{\log \left(\frac{2Dc}{\delta} \right) + 1} \right]. \end{aligned}$$

Recall that $\boldsymbol{\theta}_{[n]}$ is the closest point to $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}_n$ in terms of the L_1 distance. Then we have

$$(A.18) \quad \begin{aligned} |f(\boldsymbol{\theta}) - \mu_{n-1}(\boldsymbol{\theta}_{[n]})| &= |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_{[n]}) + f(\boldsymbol{\theta}_{[n]}) - \mu_{n-1}(\boldsymbol{\theta}_{[n]})| \\ &\leq |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_{[n]})| + |f(\boldsymbol{\theta}_{[n]}) - \mu_{n-1}(\boldsymbol{\theta}_{[n]})| \\ &\leq \frac{1}{n^2} + \sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_{[n]}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \quad \forall n \geq 1, \end{aligned}$$

with probability $\geq 1 - \delta$, where the last inequality uses (A.13) and (A.16).

Likewise, with the same β_n , we have

$$(A.19) \quad |f(\boldsymbol{\theta}) - \hat{\mu}_{n-1}(\boldsymbol{\theta}_{[n]})| \leq \frac{1}{n^2} + \sqrt{\beta_n} \hat{\sigma}_{n-1}(\boldsymbol{\theta}_{[n]}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \quad \forall n \geq 1,$$

with probability $\geq 1 - \delta$, where the inequality uses (A.13) and (A.15).

Let $\bar{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_n^I, \boldsymbol{\theta}_0^{NI})$ be the point where the portion for $k \in \mathcal{K}^I$ is from $\hat{\boldsymbol{\theta}}_n$ generated by Algorithm 2 and the portion for $k \in \mathcal{K}^{NI}$ is fixed at some $\boldsymbol{\theta}_0^{NI}$. Using Lemma A.1 with $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_n$, we have

$$(A.20) \quad \begin{aligned} &|f(\bar{\boldsymbol{\theta}}_n) - f(\hat{\boldsymbol{\theta}}_n)| \\ &\leq \alpha_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \|\hat{\boldsymbol{\theta}}_n^I - \hat{\boldsymbol{\theta}}_n^I\|_1 + \rho_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \|\boldsymbol{\theta}_0^{NI} - \hat{\boldsymbol{\theta}}_n^{NI}\|_1 \\ &= \rho_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} \|\boldsymbol{\theta}_0^{NI} - \hat{\boldsymbol{\theta}}_n^{NI}\|_1 \\ &\leq \rho_{\max} b \sqrt{\log \left(\frac{2Dc}{\delta} \right)} (D - d_I), \end{aligned}$$

with probability $\geq 1 - \delta$, where the last inequality holds due to $\boldsymbol{\Theta} := [0, 1]^D$.

Let $\bar{\boldsymbol{\theta}}_{n,[n]}$ denote the closest point in $\bar{\boldsymbol{\theta}}_n$ in $\boldsymbol{\theta}_n$ with the fixed minor parameter at $\boldsymbol{\theta}_0^{NI}$. Thus, we have $\bar{\boldsymbol{\theta}}_{n,[n]} = (\hat{\boldsymbol{\theta}}_{n,[n]}^I, \boldsymbol{\theta}_0^{NI})$ for some $\hat{\boldsymbol{\theta}}_{n,[n]}^I$. In Algorithm 1, since $\boldsymbol{\theta}_n^I$ is the minimizer of LCB with fixed $D - d_I$ parameters as $\boldsymbol{\theta}_0^{NI}$, we have

$$(A.21) \quad \mu_{n-1}(\bar{\boldsymbol{\theta}}_{n,[n]}) - \sqrt{\beta_n} \sigma_{n-1}(\bar{\boldsymbol{\theta}}_{n,[n]}) \geq \mu_{n-1}(\boldsymbol{\theta}_n) - \sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_n).$$

Similarly, in Algorithm 2, let $\theta_{[n]}^*$ denote the closest point to a true minimizer θ^* in Θ_n in the GP-LCB algorithm. Then we have

$$(A.22) \quad \hat{\mu}_{n-1}(\theta_{[n]}^*) - \sqrt{\beta_n} \hat{\sigma}_{n-1}(\theta_{[n]}^*) \geq \hat{\mu}_{n-1}(\hat{\theta}_n) - \sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\theta}_n),$$

because $\hat{\theta}_n$ is the minimizer of LCB at the n th round, representing the minimum value of LCB in the right hand side.

We introduce another lemma from Srinivas et al. (2010) and the corresponding result for Algorithm 1 to proceed further.

LEMMA A.3 (Lemma 5.5 in Srinivas et al. (2010)). *Let $\{\hat{\theta}_n\}$ be a sequence of points generated by Algorithm 2. Pick $\delta \in (0, 1)$ and set $\beta_n = 2 \log(\omega_n/\delta)$, where $\sum_{n \geq 1} \omega_n^{-1} = 1$ and $\omega_n > 0$. Then we have*

$$(A.23) \quad |f(\hat{\theta}_n) - \hat{\mu}_{n-1}(\hat{\theta}_n)| \leq \sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\theta}_n), \quad \forall n \geq 1,$$

with probability $\geq 1 - \delta$.

Extending this result, let $\{\theta_n\}$ denote a sequence of points generated by Algorithm 1. Then we have

$$(A.24) \quad |f(\theta_n) - \mu_{n-1}(\theta_n)| \leq \sqrt{\beta_n} \sigma_{n-1}(\theta_n), \quad \forall n \geq 1,$$

with probability $\geq 1 - \delta$.

On the other hand, by substituting δ with $\delta/4$ in (A.17) and setting $\omega_n = \pi^2 n^2/6$, we have

$$(A.25) \quad \begin{aligned} \beta_n = 2 \log \left(\frac{2\pi^2 n^2}{3\delta} \right) + 2d_I \log \left[\alpha_{\max} D n^2 b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} \right] \\ + 2(D - d_I) \log \left[\rho_{\max} D n^2 b \sqrt{\log \left(\frac{8Dc}{\delta} \right) + 1} \right]. \end{aligned}$$

Now, we can derive the upper bound for the instantaneous regret r_n , i.e., $r_n = f(\theta_n) - f(\theta^*) = f(\theta_n^I, \theta_0^{NI}) - f(\theta^*)$ for the simplified Ex-Sliced-Calib in Algorithm 1.

$$(A.26) \quad \begin{aligned} r_n &= f(\theta_n) - f(\theta^*) \\ &\leq f(\theta_n) - \hat{\mu}_{n-1}(\theta_{[n]}^*) + \sqrt{\beta_n} \hat{\sigma}_{n-1}(\theta_{[n]}^*) + \frac{1}{n^2} \\ &\leq f(\theta_n) - \hat{\mu}_{n-1}(\hat{\theta}_n) + \sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\theta}_n) + \frac{1}{n^2} \\ &= \underbrace{-\hat{\mu}_{n-1}(\hat{\theta}_n) + \sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\theta}_n) + f(\hat{\theta}_n)}_{(a)} \underbrace{-f(\hat{\theta}_n) + f(\bar{\theta}_n)}_{(b)} \underbrace{-f(\bar{\theta}_n) + f(\theta_n)}_{(c)} + \frac{1}{n^2} \end{aligned}$$

with probability $\geq 1 - \delta$, where the first inequality is due to (A.19) and the second inequality is due to (A.22). Let us break down the last expression into three terms (a), (b), and (c). Then we have the following.

$$(A.27) \quad (a) := -\hat{\mu}_{n-1}(\hat{\theta}_n) + \sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\theta}_n) + f(\hat{\theta}_n) \leq 2\sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\theta}_n),$$

where the inequality is satisfied because of (A.23).

$$(A.28) \quad (b) := -f(\hat{\theta}_n) + f(\bar{\theta}_n) \leq \rho_{\max} b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} (D - d_I),$$

where the inequality is satisfied because of (A.20).

$$\begin{aligned}
 (c) &:= -f(\bar{\boldsymbol{\theta}}_n) + f(\boldsymbol{\theta}_n) + \frac{1}{n^2} \leq -\mu_{n-1}(\bar{\boldsymbol{\theta}}_{n,[n]}) + \sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_{n,[n]}) + f(\boldsymbol{\theta}_n) + \frac{2}{n^2} \\
 (A.29) \quad &\leq -\mu_{n-1}(\boldsymbol{\theta}_n) + \sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_n) + f(\boldsymbol{\theta}_n) + \frac{2}{n^2} \\
 &\leq 2\sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_n) + \frac{2}{n^2},
 \end{aligned}$$

where the first, second, and last inequalities are satisfied because of (A.18), (A.21), and (A.24), respectively.

Hence, (A.26) can be upper bounded by combining the inequalities (A.27), (A.28), and (A.29) as follows.

$$(A.30) \quad r_n \leq 2\sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\boldsymbol{\theta}}_n) + \rho_{\max} b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} (D - d_I) + 2\sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_n) + \frac{2}{n^2},$$

By taking the summation from 1 to N on both sides in (A.30), we have (A.31)

$$\begin{aligned}
 R_N &= \sum_{n=1}^N r_n = \sum_{n=1}^N [f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}^*)] \\
 &= \sum_{n=1}^N 2\sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\boldsymbol{\theta}}_n) + \sum_{n=1}^N 2\sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_n) + \sum_{n=1}^N \frac{2}{n^2} + \sum_{n=1}^N \rho_{\max} b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} (D - d_I) \\
 &\leq 2\sqrt{C_1 N \beta_N \gamma_N} + \frac{\pi^2}{3} + \rho_{\max} b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} N(D - d_I),
 \end{aligned}$$

with probability $\geq 1 - \delta$, where the two terms $\sum_{n=1}^N 2\sqrt{\beta_n} \hat{\sigma}_{n-1}(\hat{\boldsymbol{\theta}}_n)$ and $\sum_{n=1}^N 2\sqrt{\beta_n} \sigma_{n-1}(\boldsymbol{\theta}_n)$ are upper bounded by $\sqrt{C_1 N \beta_N \gamma_N}$ using Lemma 5.4 in Srinivas et al. (2010) and the Cauchy-Schwarz inequality with $C_1 = 8/\log(1 + \tilde{\sigma}^{-2})$, where $\tilde{\sigma}^2$ is the nugget when we fit GP for $\{\boldsymbol{\theta}, \tilde{f}(\boldsymbol{\theta})\}_{\boldsymbol{\theta} \in A}$ and $\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + e$ with $e \sim \mathcal{N}(0, \tilde{\sigma}^2)$. Here, γ_N is the maximum information gain with a finite set of N sampling points, that is,

$$(A.32) \quad \gamma_N = \max_{A \subset \boldsymbol{\Theta}: |A|=N} \frac{1}{2} \log \left(\det(I + \tilde{\sigma}^2 K_A) \right).$$

A similar approach that provides the bounds on the cumulative regret under the noise-free setting in BO can be found in Lyu, Yuan and Tsang (2019).

Finally, from (A.31), we have the averaged regret as

$$(A.33) \quad \frac{R_N}{N} \leq 2\sqrt{C_1 \frac{\beta_N \gamma_N}{N}} + \frac{\pi^2}{3N} + \rho_{\max} b \sqrt{\log \left(\frac{8Dc}{\delta} \right)} (D - d_I),$$

with probability $\geq 1 - \delta$. □

B. Estimation of the Fisher Information Matrix. Let us consider the log-likelihood function from a single observation with real-valued $y(\cdot)$ and $\eta(\cdot)$ as follows.

$$(A.34) \quad \ell_1(\boldsymbol{\theta}^I | y(\mathbf{x}), \boldsymbol{\theta}^{NI}) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y(\mathbf{x}) - \eta(\mathbf{x}; \boldsymbol{\theta}^I, \boldsymbol{\theta}^{NI}))^2.$$

Consider the expected Fisher information matrix $\mathcal{I}(\boldsymbol{\theta}_0^I)$ expressed as

$$(A.35) \quad \mathcal{I}(\boldsymbol{\theta}_0^I) = E \left[\frac{\partial \ell_1(\boldsymbol{\theta}^I | y(\mathbf{x}), \boldsymbol{\theta}^{NI})}{\partial(\boldsymbol{\theta}^I)} \right] \left[\frac{\partial \ell_1(\boldsymbol{\theta}^I | y(\mathbf{x}), \boldsymbol{\theta}^{NI})}{\partial(\boldsymbol{\theta}^I)} \right]^T \Big|_{\boldsymbol{\theta}^I = \boldsymbol{\theta}_0^I}.$$

Then this expected Fisher information matrix can be approximated by the following observed Fisher information matrix.

$$(A.36) \quad \mathcal{I}(\hat{\boldsymbol{\theta}}_{\text{ML}}^I) \approx \frac{1}{N_x} \sum_{i=1}^{N_x} \left[\frac{\partial \ell_1(\boldsymbol{\theta}^I | y(\mathbf{x}_i), \boldsymbol{\theta}^{NI})}{\partial(\boldsymbol{\theta}^I)} \right] \left[\frac{\partial \ell_1(\boldsymbol{\theta}^I | y(\mathbf{x}_i), \boldsymbol{\theta}^{NI})}{\partial(\boldsymbol{\theta}^I)} \right]^T \Big|_{\boldsymbol{\theta}^I = \hat{\boldsymbol{\theta}}_{\text{ML}}^I}.$$

Since the first-order partial derivative of $\eta(\mathbf{x}_i; \boldsymbol{\theta}^I, \boldsymbol{\theta}^{NI})$ is not available, we obtain them from the numerical approximation using central finite-differencing (Abramowitz and Stegun, 1972).

$$(A.37) \quad \frac{\partial \ell_1(\boldsymbol{\theta}^I | y(\mathbf{x}), \boldsymbol{\theta}^{NI})}{\partial \theta_k} \approx \frac{\partial \ell_1(\boldsymbol{\theta}^I + h \mathbf{e}_k | y(\mathbf{x}), \boldsymbol{\theta}^{NI}) - \partial \ell_1(\boldsymbol{\theta}^I - h \mathbf{e}_k | y(\mathbf{x}), \boldsymbol{\theta}^{NI})}{2h},$$

for all $k \in \mathcal{K}^I$, where $h > 0$ denotes a small perturbation (e.g., 10^{-8}), and $\mathbf{e}_k$ is the k th standard basis vector in $\mathbb{R}^{d_I}$, i.e., a column vector of length d_I with a one in the k th position and zeros elsewhere.

C. Sensitivity Analysis and Guideline for Choosing Algorithmic Parameters. We examine the influence of various algorithmic parameters, namely, d , p , and σ_0 , on the performance of Ex-Sliced-Calib.

C1. Effect of d : Table 1 and Figure 1 show the results with different d values. Ex-Sliced-Calib generally exhibits improved performance in terms of calibration accuracy (i.e., smaller MSEs) and stability (i.e., smaller standard deviations) when the trial dimension d is sufficiently small. We believe that reducing the trial dimension contributes to improved calibration by curtailing excessive exploration, which can be particularly prevalent in high-dimensional spaces. Simultaneously, it enables the calibration process to concentrate on those parameters that are truly significant, thereby enhancing the efficiency and effectiveness of the search. Based on these results, for $D = 20 \sim 30$, $d = \lceil D/3 \rceil - \lfloor D/2 \rfloor$ that covers 33% to 50% of the entire parameters appears to be a reasonable choice for Ex-Sliced-Calib.

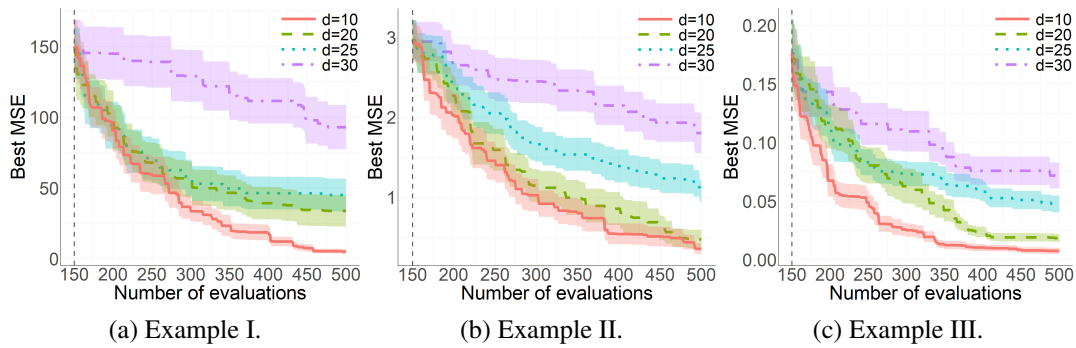


Fig 1: Trajectories of the lowest MSEs with different trial dimensions d in the numerical study (Note: the band denotes the ± 1 standard error).

TABLE 1

Sensitivity analysis with respect to d in Ex-Sliced-Calib in the numerical study (Note: the values inside parentheses are standard deviations).

Example	d	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	10	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	20	33.726 (49.748)	388.135 (163.180)	383.377 (161.011)	—
	25	44.887 (52.024)	386.275 (161.882)	380.205 (160.768)	—
	30	92.890 (69.778)	278.484 (292.822)	311.888 (266.535)	—
	True parameter values		420.969	420.969	—
II	10	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	20	0.475 (0.522)	0.023 (0.061)	-0.022 (0.084)	—
	25	1.082 (0.645)	0.028 (0.114)	-0.059 (0.240)	—
	30	1.807 (1.136)	0.018 (0.219)	0.060 (0.431)	—
	True parameter values		0	0	—
III	10	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	20	0.018 (0.015)	0.132 (0.070)	0.557 (0.015)	0.854 (0.009)
	25	0.048 (0.032)	0.196 (0.116)	0.559 (0.020)	0.850 (0.015)
	30	0.072 (0.050)	0.251 (0.191)	0.558 (0.023)	0.849 (0.016)
	True parameter values		0.115	0.556	0.853

C2. *Effect of p* : Table 2 and Figure 2 indicate that overall, assigning a positive probability to p in (8) can lead to enhanced performance in calibration accuracy in many cases. Based on the results, we suggest setting p between 0.05 and 0.1 in Ex-Sliced-Calib. This range strikes a practical balance between the advantage of increased space-filling and the need to preserve a focused search on the regions of greatest impact.

TABLE 2

Sensitivity analysis with respect to p in Ex-Sliced-Calib in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	p	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	0	9.263 (25.325)	421.422 (4.181)	419.121 (7.687)	—
	0.05	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	0.25	23.456 (39.065)	423.643 (12.176)	421.324 (6.670)	—
	0.5	24.505 (26.044)	418.977 (10.904)	417.803 (8.656)	—
	True parameter values		420.969	420.969	—
II	0	0.525 (0.605)	0.010 (0.088)	-0.003 (0.082)	—
	0.05	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	0.25	0.631 (0.457)	-0.009 (0.105)	-0.003 (0.068)	—
	0.5	0.820 (0.633)	0.048 (0.103)	0.030 (0.104)	—
	True parameter values		0	0	—
III	0	0.006 (0.007)	0.120 (0.023)	0.556 (0.008)	0.853 (0.006)
	0.05	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	0.25	0.011 (0.019)	0.129 (0.030)	0.558 (0.007)	0.856 (0.009)
	0.5	0.016 (0.021)	0.114 (0.050)	0.557 (0.011)	0.849 (0.010)
	True parameter values		0.115	0.556	0.853

C2. *Effect of σ_0* : The standard deviation σ_0 in the weight distribution for each parameter dictates the extent of exploration. With large (resp., small) σ_0 , the weights of the parameters are drawn from a wider (resp., narrower) distribution around their respective posterior means, encouraging a higher degree of exploration. Based on our observations in Table 3 and Figure 3, the calibration process tends to exhibit good performance when setting σ_0 to approximately 0.5 to 0.6.

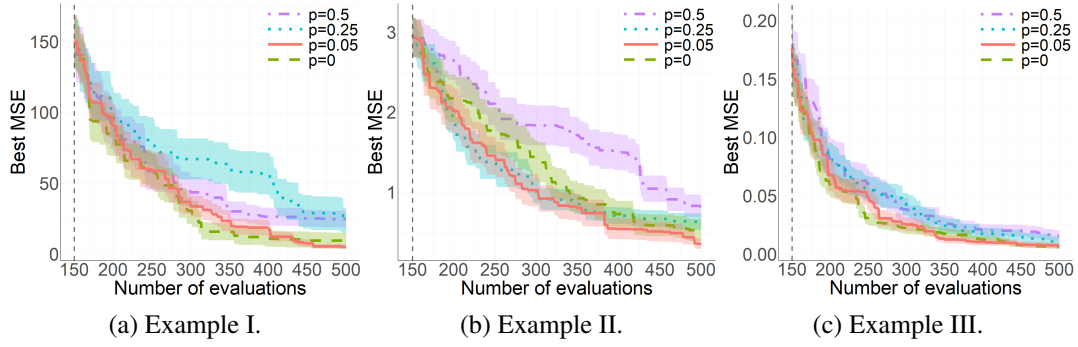


Fig 2: Trajectories of the lowest MSEs with different probabilities p in the numerical study (Note: the band denotes the ± 1 standard error).

TABLE 3

Sensitivity analysis with respect to σ_0 in Ex-Sliced-Calib in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	σ_0	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	0.4	14.407 (26.552)	420.604 (6.048)	382.790 (162.098)	—
	0.5	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	0.6	13.754 (22.579)	420.503 (8.271)	419.592 (6.803)	—
	True parameter values		420.969	420.969	—
II	0.4	0.565 (0.627)	-0.048 (0.227)	0.010 (0.068)	—
	0.5	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	0.6	0.528 (0.550)	-0.015 (0.097)	-0.023 (0.062)	—
	True parameter values		0	0	—
III	0.4	0.009 (0.010)	0.122 (0.027)	0.551 (0.008)	0.852 (0.007)
	0.5	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	0.6	0.012 (0.012)	0.118 (0.016)	0.557 (0.015)	0.851 (0.006)
	True parameter values		0.115	0.556	0.853

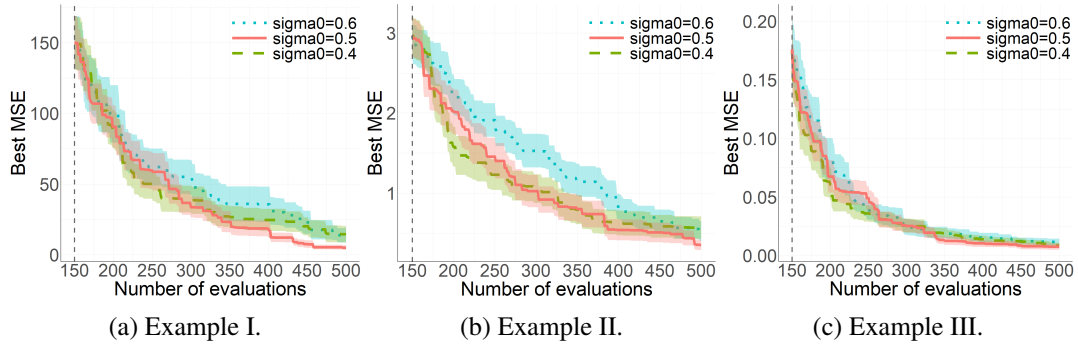


Fig 3: Trajectories of the lowest MSEs with different σ_0 in the simulation study (Note: the band denotes the ± 1 standard error).

D. Comparison of Sensitivity Indices. To check if Ex-Sliced-Calib captures the true influence correctly, we calculate total sensitivity indices without constructing a GP surrogate model using Monte Carlo methods (Saltelli, 2002) for all examples in the numerical study. We

treat these Sobol' indices as "true importance" of each parameter and compare the posterior mean obtained by Ex-Sliced-Calib at the maximum budget of $N_{\max}$.

Table 4 reports the true and estimated values of importance measures of all parameters. We note that Ex-Sliced-Calib captures the true degree of sensitivity for all influential parameters; In Examples I and II, θ_1 and θ_2 are influential parameters, in Example III, θ_2 and θ_3 are the most important parameters, followed by θ_1 . For the non-influential parameters in Examples I and II, the posterior means are larger than zeros. However, the posterior means of unimportant parameters are still significantly lower than those of influential parameters, indicating that Ex-Sliced-Calib can discern influential parameters with non-influential ones.

TABLE 4

Comparison of the final posterior means of each weight distribution from Ex-Sliced-Calib and the "true" importance measures via Sobol' indices (Note: the values inside parentheses are standard deviations, and the bold numbers indicate the posterior means of the influential parameters).

Parameter	Example I		Example II		Example III	
	Ex-Sliced-Calib	Sobol	Ex-Sliced-Calib	Sobol	Ex-Sliced-Calib	Sobol
θ_1	0.54 (0.08)	0.50 (0.03)	0.48 (0.05)	0.54 (0.04)	0.05 (0.03)	0.08 (0.01)
θ_2	0.50 (0.05)	0.51 (0.04)	0.50 (0.10)	0.54 (0.04)	0.44 (0.03)	0.45 (0.03)
θ_3	0.38 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.86 (0.04)	0.84 (0.04)
θ_4	0.38 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.04 (0.02)	0.00 (0.00)
θ_5	0.38 (0.09)	0.00 (0.00)	0.31 (0.14)	0.00 (0.00)	0.03 (0.04)	0.00 (0.00)
θ_6	0.37 (0.08)	0.00 (0.00)	0.29 (0.12)	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
θ_7	0.37 (0.08)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.03 (0.02)	0.00 (0.00)
θ_8	0.38 (0.09)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.02 (0.03)	0.00 (0.00)
θ_9	0.38 (0.08)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)
θ_{10}	0.39 (0.09)	0.00 (0.00)	0.29 (0.12)	0.00 (0.00)	0.04 (0.02)	0.00 (0.00)
θ_{11}	0.38 (0.08)	0.00 (0.00)	0.30 (0.12)	0.00 (0.00)	0.02 (0.03)	0.00 (0.00)
θ_{12}	0.38 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
θ_{13}	0.38 (0.09)	0.00 (0.00)	0.30 (0.12)	0.00 (0.00)	0.02 (0.03)	0.00 (0.00)
θ_{14}	0.39 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.03 (0.02)	0.00 (0.00)
θ_{15}	0.38 (0.09)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.03 (0.04)	0.00 (0.00)
θ_{16}	0.38 (0.10)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.02 (0.03)	0.00 (0.00)
θ_{17}	0.38 (0.08)	0.00 (0.00)	0.31 (0.11)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)
θ_{18}	0.39 (0.09)	0.00 (0.00)	0.30 (0.13)	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
θ_{19}	0.38 (0.10)	0.00 (0.00)	0.31 (0.11)	0.00 (0.00)	0.03 (0.02)	0.00 (0.00)
θ_{20}	0.37 (0.09)	0.00 (0.00)	0.31 (0.11)	0.00 (0.00)	0.04 (0.03)	0.00 (0.00)
θ_{21}	0.38 (0.09)	0.00 (0.00)	0.30 (0.13)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)
θ_{22}	0.38 (0.09)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)
θ_{23}	0.39 (0.08)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.03 (0.04)	0.00 (0.00)
θ_{24}	0.38 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.02 (0.03)	0.00 (0.00)
θ_{25}	0.39 (0.10)	0.00 (0.00)	0.32 (0.13)	0.00 (0.00)	0.04 (0.03)	0.00 (0.00)
θ_{26}	0.38 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)
θ_{27}	0.37 (0.08)	0.00 (0.00)	0.31 (0.12)	0.00 (0.00)	0.02 (0.03)	0.00 (0.00)
θ_{28}	0.38 (0.08)	0.00 (0.00)	0.31 (0.10)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)
θ_{29}	0.37 (0.08)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
θ_{30}	0.38 (0.09)	0.00 (0.00)	0.30 (0.11)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)

E. Comparison of Computational Time. Overall, Ex-Sliced-Calib achieves notable computational savings relative to Full-Calib and TR-Calib through dimensionality reduction, though it requires more time than Rand-Sliced-Calib. This is because Ex-Sliced-Calib, introduces an additional step of constructing a surrogate model in the full parameter space to compute the sensitivity indices via the FAST method. While this step does add some computational overhead, we emphasize the following aspects of Ex-Sliced-Calib.

- **Computational Impact of Surrogate:** The surrogate used for computing sensitivity indices is constructed using existing data already collected in the calibration process, so it does not require any additional simulator evaluations. Although constructing the surrogate and performing the Fourier-based analysis incur some computational cost, this cost may be negligible compared with the cost of a simulation run using the (moderately) expensive simulator employed in this study and in many applications.
- **Efficiency Gains through Dimensionality Reduction:** The additional computational effort is offset by the significant efficiency gained through strategic dimensionality reduction. By probabilistically selecting influential parameters using the sensitivity-informed weight distributions, our method avoids wasting simulations on unimportant or redundant parameter combinations. This leads to better surrogate model quality in both reduced and full parameter space and faster convergence in fewer iterations.

We compare the computation time of Ex-Sliced-Calib with Rand-Sliced-Calib, Full-Calib, and TR-Calib across three numerical examples, as shown in Table 5. For the experiments, we use 20 logical processors to run 20 experiments in parallel, with each experiment occupying one logical processor. While Ex-Sliced-Calib takes more time than Rand-Sliced-Calib due to the additional surrogate modeling, it still provides a substantial reduction in computation time compared to Full-Calib and TR-Calib. In particular, under parallel execution, Ex-Sliced-Calib requires fewer overall resources than Full-Calib and TR-Calib while achieving much better calibration accuracy.

TABLE 5

Comparison of the computation times from 20 experiments in the numerical study. (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font. “Avg. time/processor” denotes the average time (with standard deviation) per experiment using one logical processor. “Total (sequential)” refers to the total time required if only one logical processor is used sequentially for 20 experiments, and “Total (parallel)” refers to the total time when 20 logical processors are used in parallel).

Example	Method	Computation Time (sec)		
		Avg. time/processor	Total (sequential)	Total (parallel)
I	Ex-Sliced-Calib	14151 (1917)	283020	18369
	Rand-Sliced-Calib	4775 (1046)	95496	7147
	Full-Calib	31906 (1645)	638111	34509
	TR-Calib	27177 (1806)	543540	30581
II	Ex-Sliced-Calib	16022 (1116)	320444	18231
	Rand-Sliced-Calib	7193 (690)	143867	8363
	Full-Calib	31819 (1576)	636372	34493
	TR-Calib	29605 (1163)	592102	31491
III	Ex-Sliced-Calib	25006 (1447)	500130	28325
	Rand-Sliced-Calib	8070 (524)	161398	9080
	Full-Calib	30679 (1630)	613571	32613
	TR-Calib	27755 (1806)	555092	31275

In high-dimensional and expensive simulation environments, the tradeoff between minimal overhead from the global surrogate and the efficiency gains in sequential design would be more clear. More importantly, our approach offers the added benefit of interpretability by quantifying parameter importance, which is not available in random dropout methods.

F. Effect of the Boundary Perturbation Term $h(\cdot)$. We empirically examine the impact of $h(\cdot)$ by conducting comparative experiments with and without its incorporation. Table 6 reports the test MSEs and calibrated parameter values from 20 independent runs of Ex-Sliced-Calib and Rand-Sliced-Calib, both with and without $h(\cdot)$. Figure 4 illustrates the convergence behavior of both methods when $h(\cdot)$ is excluded.

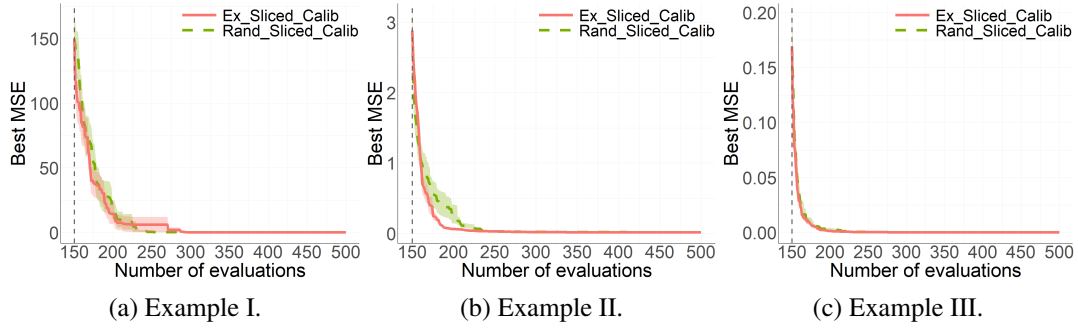


Fig 4: Trajectories of the lowest MSEs without incorporating $h(\cdot)$ in the numerical study (Note: the band denotes the ± 1 standard error).

Overall, when $h(\cdot)$ is excluded, Ex-Sliced-Calib and Rand-Sliced-Calib perform similarly, achieving higher calibration accuracy and faster convergence compared to the setting with $h(\cdot)$. Even in this case, Ex-Sliced-Calib yields smaller MSEs in Examples I and II. By contrast, when $h(\cdot)$ is included, reflecting the added difficulty of calibration tasks, Ex-Sliced-Calib significantly outperforms Rand-Sliced-Calib in terms of MSE in all cases. This comparison underscores the practical value of Ex-Sliced-Calib under model misspecification, while also demonstrating that it retains an advantage over Rand-Sliced-Calib even in the absence of misspecification.

TABLE 6

Comparison of test MSEs and calibrated parameter values from 20 experiments with and without incorporating $h(\cdot)$ in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	$h(\cdot)$	Method	Test MSE f	Calibrated Parameter Values		
				θ_1	θ_2	θ_3
I	Yes	Ex-Sliced-Calib	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
		Rand-Sliced-Calib	13.935 (36.746)	420.308 (2.758)	348.473 (223.151)	—
	No	Ex-Sliced-Calib	0.042 (0.030)	420.978 (0.337)	420.849 (0.374)	—
		Rand-Sliced-Calib	0.051 (0.056)	420.902 (0.411)	420.899 (0.396)	—
	True parameter values			420.969	420.969	—
II	Yes	Ex-Sliced-Calib	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
		Rand-Sliced-Calib	0.562 (0.577)	0.006 (0.065)	0.024 (0.101)	—
	No	Ex-Sliced-Calib	0.014 (0.003)	0.000 (0.001)	0.000 (0.001)	—
		Rand-Sliced-Calib	0.016 (0.004)	0.000 (0.002)	0.000 (0.002)	—
	True parameter values			0	0	—
III	Yes	Ex-Sliced-Calib	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
		Rand-Sliced-Calib	0.010 (0.031)	0.116 (0.016)	0.555 (0.008)	0.850 (0.009)
	No	Ex-Sliced-Calib	0.000 (0.000)	0.114 (0.004)	0.556 (0.001)	0.853 (0.000)
		Rand-Sliced-Calib	0.000 (0.000)	0.114 (0.005)	0.556 (0.000)	0.852 (0.000)
	True parameter values			0.115	0.556	0.853

G. Effect of Parameter Sparsity Levels. To evaluate the performance of the proposed Ex-Sliced-Calib method under different sparsity levels, we conduct numerical experiments in which different numbers of influential parameters, d_I , were set to be the true influential ones out of the total $D = 30$ parameters: $d_I = 2, 10, 20$ and 30 . The last scenario corresponds to the case where all calibration parameters affect the model outputs (i.e., no sparsity). Table 7 reports the test MSEs from 20 experiments across the different sparsity levels with 500 evaluation budgets for Examples I and II. Note that experiments for the different number of d_I

were conducted only for Examples I and II, since Example III cannot be extended to higher parameter spaces, as its dimension is fixed at three.

TABLE 7

Comparison of test MSEs from 20 experiments for the different number of d_I with $D = 30$ in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	Effect Sparsity	Test MSE f			
		$d_I = 2$	$d_I = 10$	$d_I = 20$	$d_I = 30$
I	Ex-Sliced-Calib	4.976 (7.405)	1080.516 (271.695)	3868.781 (496.902)	7142.532 (537.283)
	Rand-Sliced-Calib	13.935 (36.746)	1107.310 (249.390)	3809.455 (431.844)	6945.683 (510.778)
	Full-Calib	92.89 (69.778)	2516.637 (217.064)	5954.514 (334.313)	9664.795 (319.774)
	TR-Calib	71.931 (70.363)	1383.670 (214.988)	3406.570 (439.657)	5415.124 (617.103)
II	Ex-Sliced-Calib	0.355 (0.295)	4.868 (1.183)	7.878 (0.597)	9.147 (0.352)
	Rand-Sliced-Calib	0.562 (0.577)	4.738 (0.690)	7.353 (0.913)	8.860 (0.450)
	Full-Calib	1.807 (1.136)	8.113 (0.841)	9.640 (0.540)	10.103 (0.517)
	TR-Calib	0.783 (1.153)	4.711 (1.585)	6.653 (0.993)	7.819 (1.051)

When d_I is 2 or 10, which are less than 30% of the total parameter set, our approach outperforms other methods. When d_I is close to 20 (67% of the total parameter set), its performance is comparable to that of the best-performing method, TR-Calib. However, for $d_I = 30$, our approach underperforms relative to TR-Calib. This outcome for large d_I is expected, as Ex-Sliced-Calib’s trial dimension may not contain all the influential parameters at once, preventing it from fully capturing dependencies among them when designing the next query point.

In our case study with the BEM application, three to four out of thirty parameters turn out to be significant, accounting for about 10%–13%. These results from both numerical examples and case study indicate that while the superiority of our approach is most pronounced for small d_I , it remains competitive even when d_I is moderately large.

H. Choice of an Acquisition Function. To evaluate the robustness of our method to the choice of acquisition function, we conduct experiments comparing Ex-Sliced-Calib under four widely used acquisition functions: Expected Improvement (EI) (Jones, Schonlau and Welch, 1998), Lower Confidence Bound (LCB) (Srinivas et al., 2010), Probability of Improvement (PI) (Kushner, 1964), and Knowledge Gradient (KG) (Frazier, Powell and Dayanik, 2009). The following provides a brief introduction to each acquisition function.

- **EI:** It selects the next query point by maximizing the expected amount by which the objective function will improve over the current best observation. Let $f_{\min}$ be the best (i.e., lowest) observed function value so far, and suppose the predictive distribution of the surrogate model at point θ is Gaussian: $f(\theta) \sim \mathcal{N}(\mu(\theta), \sigma^2(\theta))$. Then, EI is given by

$$(A.38) \quad \mathcal{A}_{\text{EI}}(\theta) = E[\max(f_{\min} - f(\theta), 0)] = (f_{\min} - \mu(\theta))\Phi(z) + \sigma(\theta)\phi(z),$$

where $z = (f_{\min} - \mu(\theta))/\sigma(\theta)$, and $\Phi(\cdot)$, $\phi(\cdot)$ denote the CDF and PDF of the standard normal distribution, respectively. EI naturally balances exploration (through large $\sigma(\theta)$) and exploitation (through small $\mu(\theta)$).

- **LCB:** It selects the next query point based on a linear combination of the predictive mean $\mu(\theta)$ and standard deviation $\sigma(\theta)$. For minimization problems, the LCB acquisition function is minimized to determine the next design, and it is given by

$$(A.39) \quad \mathcal{A}_{\text{LCB}}(\theta) = \mu(\theta) - \kappa\sigma(\theta),$$

where $\kappa > 0$ controls the degree of exploration. Larger values of κ prioritize exploration by emphasizing uncertainty, while smaller values emphasize exploitation of the predictive mean.

- **PI:** It chooses the next query point by maximizing the probability that the function value will improve upon the current best value. It is defined as

$$(A.40) \quad \mathcal{A}_{PI}(\theta) = P(f(\theta) < f_{\min} - \xi) = \Phi \left(\frac{f_{\min} - \mu(\theta) - \xi}{\sigma(\theta)} \right),$$

where $\xi > 0$ is a trade-off parameter to encourage exploration. PI is simple and interpretable, but can overly favor regions with low predictive mean even if the uncertainty is low.

- **KG:** It selects the next query point by maximizing the the expected increase in the best posterior mean after evaluating the function at a candidate point. Let $\mu_n(\theta')$ be the current GP posterior mean and $\mu_{n+1}^{\theta}(\theta')$ the updated posterior mean after querying θ . KG is defined as

$$(A.41) \quad \mathcal{A}_{KG}(\theta) = E \left[\max_{\theta'} \mu_{n+1}^{\theta}(\theta') \right] - \max_{\theta'} \mu_n(\theta'),$$

which represents the expected benefit of learning from sampling at θ . KG is particularly useful in scenarios with expensive queries and limited evaluation budgets.

TABLE 8

Comparison of test MSEs and calibrated parameter values from 20 experiments using Ex-Sliced-Calib under different acquisition functions in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	Acquisition Function $\mathcal{A}$	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	EI	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	LCB	42.521 (62.162)	291.039 (312.858)	401.687 (244.079)	—
	PI	7.042 (7.591)	419.856 (4.857)	421.280 (5.828)	—
	KG	62.977 (52.774)	-319.415 (849.897)	260.581 (626.321)	—
	True parameter values		420.969	420.969	—
II	EI	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	LCB	0.501 (0.500)	-0.020 (0.068)	0.019 (0.082)	—
	PI	0.246 (0.278)	0.006 (0.035)	-0.004 (0.056)	—
	KG	1.468 (0.821)	-0.083 (0.259)	-0.055 (0.132)	—
	True parameter values		0	0	—
III	EI	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	LCB	0.010 (0.010)	0.119 (0.033)	0.556 (0.009)	0.850 (0.008)
	PI	0.006 (0.008)	0.115 (0.026)	0.556 (0.009)	0.854 (0.006)
	KG	0.054 (0.053)	0.117 (0.051)	0.548 (0.026)	0.850 (0.018)
	True parameter values		0.115	0.556	0.853

Table 8 summarizes the performance of Ex-Sliced-Calib under different acquisition functions across three numerical examples. The results show that EI consistently provides the best or near-best performance, whereas LCB and KG yield inferior performance in these challenging settings, possibly due to their tendency toward excessive exploration. Further, PI performs competitively, especially in the smoother setting in Example III.

Considering that the convergence result in the theoretical regret bound analysis was derived under LCB, the superior empirical performance observed under EI suggests that a similar or possibly tighter regret bound can be established.

I. Validation of the Model Assumptions. There are two GPs used in the proposed Ex-Sliced-Calib method. One is the GP used as a surrogate for global sensitivity analysis in the original parameter space, and the other is the GP that guides the sequential design within BO in the reduced space.

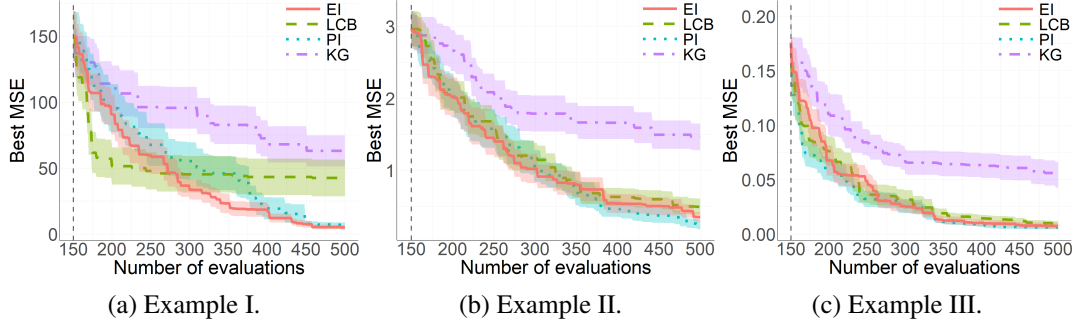


Fig 5: Trajectories of the lowest MSEs using Ex-Sliced-Calib under different acquisition functions in the numerical study (Note: the band denotes the ± 1 standard error).

The GP framework assumes that the joint distribution of function values (i.e., loss function values) at any finite set of inputs (i.e., calibration parameters) is multivariate Gaussian. Furthermore, the underlying function is assumed to be continuous and reasonably smooth, with smoothness encoded via a kernel function, such as the squared exponential or Matérn kernel. GPs also assume that the observation noise is independently distributed Gaussian.

First, it should be noted that we use GPs as surrogate models to map the parameter space to loss function values. The GPs define a prior over functions rather than over output distributions. Thus, GPs do not require the observational data itself (e.g., electricity consumption in the BEM case study) to be Gaussian-distributed.

Having said that, we conducted additional simulation studies in which the ground truth data exhibit non-Gaussian characteristics, including (i) heavy-tailed noise (e.g., Student- t distribution), (ii) uniform noise, and (iii) skewed noise (e.g., skewed-Gaussian distribution). Specifically, we compare our numerical experiment setting, Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with (i) Student- t noise $\epsilon \sim t(df) \times \sigma$ where the degrees of freedom $df = 4$, (ii) uniform noise $\epsilon \sim \text{Uniform}(-2\sigma, 2\sigma)$, and (iii) skewed-Gaussian noise $\epsilon \sim \mathcal{SN}(0, \sigma^2, \alpha_{\text{shape}})$ where the shape parameter $\alpha_{\text{shape}} = 5$. For Examples I and II, we set $\sigma = 0.1$, and for Example III, we use $\sigma = 0.01$. Figure 6 illustrates the noise distributions used to evaluate the robustness of Ex-Sliced-Calib.

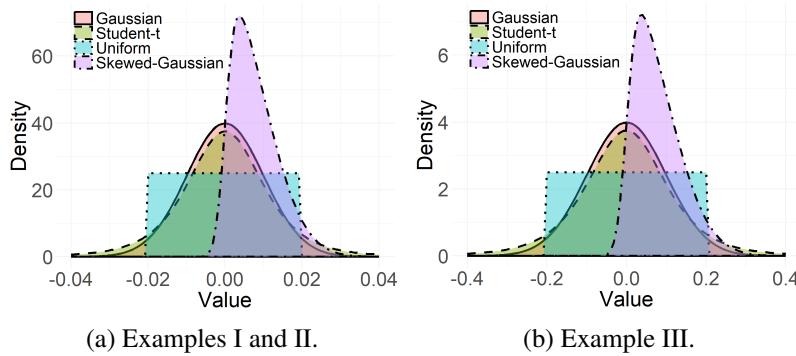


Fig 6: Different noise distributions used to assess the robustness of Ex-Sliced-Calib.

Overall, under Gaussian noise, which perfectly aligns with the GP assumption, Ex-Sliced-Calib most effectively guides the search toward higher-quality parameter values compared to

other methods, as shown in Table 9 and Figure 7. On the other hand, calibration performance slightly deteriorates under non-Gaussian noise, resulting in higher MSEs and standard deviations. However, the calibrated values for non-Gaussian cases are close to those for the Gaussian case; see the last three columns of Table 9. It suggests that although non-Gaussian data slightly worsens performance, the proposed method remains robust. Thus, the GP surrogates used for both global sensitivity analysis and sequential design in BO remain effective for non-Gaussian cases. The robustness of GPs to moderate violations of the Gaussian assumption has been discussed in the literature, provided that the underlying function is sufficiently smooth and the noise is not highly structured or adversarial (Rasmussen and Williams, 2006; Gramacy, 2020).

TABLE 9

Comparison of test MSEs and calibrated parameter values from 20 experiments using Ex-Sliced-Calib under different noise assumptions (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	Noise Distribution	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	Gaussian	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	Student- t	12.138 (17.935)	421.560 (6.040)	422.545 (7.903)	—
	Uniform	12.934 (17.945)	419.959 (8.657)	421.610 (5.704)	—
	Skewed-Gaussian	17.845 (26.456)	422.026 (10.282)	421.725 (6.307)	—
	True parameter values		420.969	420.969	—
II	Gaussian	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	Student- t	0.641 (0.549)	-0.007 (0.066)	-0.011 (0.111)	—
	Uniform	0.608 (0.634)	-0.071 (0.223)	-0.006 (0.067)	—
	Skewed-Gaussian	0.561 (0.567)	-0.008 (0.068)	0.000 (0.084)	—
	True parameter values		0	0	—
III	Gaussian	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	Student- t	0.007 (0.011)	0.115 (0.029)	0.553 (0.007)	0.853 (0.007)
	Uniform	0.008 (0.008)	0.105 (0.028)	0.557 (0.009)	0.853 (0.007)
	Skewed-Gaussian	0.006 (0.005)	0.115 (0.021)	0.556 (0.010)	0.852 (0.003)
	True parameter values		0.115	0.556	0.853

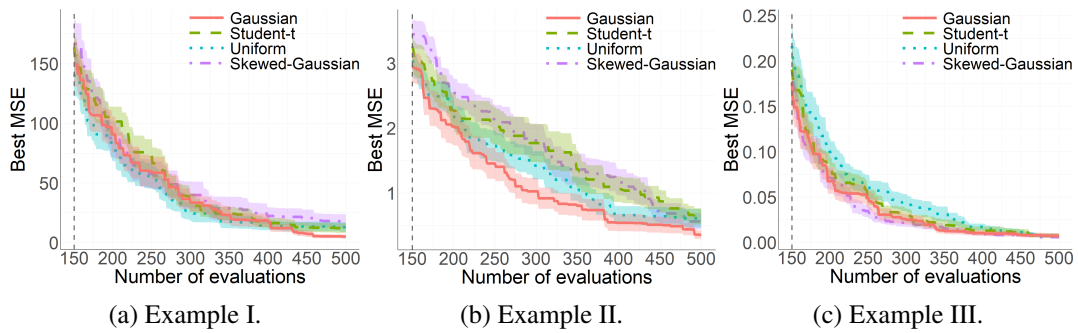


Fig 7: Trajectories of the lowest MSEs using Ex-Sliced-Calib under different noise assumptions in the numerical study (Note: the band denotes the ± 1 standard error).

Next, regarding the continuity and smoothness assumption, the loss functions in the numerical study are continuous and smooth. In the BEM case study, the dataset includes physical observations and computer model outputs generated via EnergyPlus. Since the response

surface of the simulation, relating inputs such as weather conditions and calibration parameters to energy usage, is likely to be smooth and continuous, and our loss function is based on squared differences using the ℓ_2 loss, the GP surrogates can be appropriately used to approximate this smooth scalar loss surface.

Additionally, in this study, we use an anisotropic kernel for both GP surrogates, commonly referred to as the Automatic Relevance Determination (ARD) kernel (Rasmussen and Williams, 2006; Neal, 2012). Specifically, we employ the anisotropic Matérn 5/2 kernel, although anisotropic versions of the squared exponential and Matérn 3/2 kernels can also be used. This choice allows each dimension of the calibration parameters to have its own length-scale, which determines how sensitive the surrogate model is to variation along that specific direction. We adopt this anisotropic kernel rather than an isotropic one for two main reasons.

- **Capturing Heterogeneous Parameter Effects:** In the simulation and BEM case studies, the calibration parameters exhibit highly varying influence on the computer model output. Some parameters significantly affect the model output (e.g., solar transmittance or lighting level), while others have minimal impact. Particularly for the GP used in global sensitivity analysis, an anisotropic kernel enables it to capture different levels of relevance across dimensions, allowing it to adapt more flexibly to this heterogeneity. This capability is essential for identifying influential parameters in high-dimensional and effect-imbalanced problems like those considered in our study.
- **Adaptability to the Isotropic Case:** Even if the underlying loss function is truly isotropic, the anisotropic (or ARD) kernel can still perform well by converging to equal lengthscales across all dimensions when fitting the GP model. Thus, the anisotropic kernel is a more general choice. That is, it performs well in both anisotropic and isotropic scenarios, while an isotropic kernel fails to capture directional effects when they exist.

As supported by the literature (Rasmussen and Williams, 2006; Gramacy, 2020), isotropic kernels can underperform in settings where parameter effects vary significantly across dimensions, potentially reducing surrogate model fidelity and leading to suboptimal points via acquisition function optimization in BO. By contrast, the anisotropic (or ARD) kernel offers greater modeling flexibility by naturally capturing heterogeneity when present, while remaining adaptable in isotropic cases by learning similar lengthscales across dimensions. For these reasons, we deliberately employ anisotropic kernels for both GP surrogates in our framework.

To compare the performance of the anisotropic-kernel GP, as adopted in this study, with that of the isotropic-kernel GP for the loss function surrogate, we conduct numerical studies using Ex-Sliced-Calib with both kernels. Table 10 below reports the test MSE values for isotropic-kernel GP vs. anisotropic-kernel GP, and Figure 8 below presents the corresponding trajectories of the best MSEs over rounds. The results indicate that using the anisotropic-kernel GP as the loss function surrogate in Ex-Sliced-Calib performs similarly to or better than using the isotropic-kernel GP in terms of calibration accuracy. In particular, in Examples I and II, Ex-Sliced-Calib converges faster when using the anisotropic-kernel GP. This is because the anisotropic-kernel GP surrogate (i) better captures heterogeneous parameter effects and (ii) remains adaptable even when the underlying loss function is truly isotropic, as described above. Note that in Example II, although the calibrated parameters for the isotropic case are slightly closer to zero, their higher variability leads to higher test MSE in some runs. Moreover, since Table 10 reports test MSE and calibration is performed on the training set, the test MSE can appear slightly higher than that of the anisotropic case.

Next, we evaluate the validity of using a stationary kernel (i.e., the anisotropic kernel) compared to a nonstationary kernel for the GP surrogate in Ex-Sliced-Calib, we add a comparison with a representative nonstationary GP model, namely the deformation-based GP (Sampson and Guttorp, 1992; Schmidt and O’Hagan, 2003).

TABLE 10

Comparison of test MSEs and calibrated parameter values from 20 experiments using isotropic- and anisotropic-kernel GPs in Ex-Sliced-Calib in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	Kernel	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	Isotropic	8.118 (9.318)	418.045 (5.157)	421.581 (5.674)	—
	Anisotropic	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	True parameter values		420.969	420.969	—
II	Isotropic	0.446 (0.337)	-0.003 (0.066)	0.001 (0.070)	—
	Anisotropic	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	True parameter values		0	0	—
III	Isotropic	0.008 (0.008)	0.112 (0.030)	0.560 (0.010)	0.853 (0.006)
	Anisotropic	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	True parameter values		0.115	0.556	0.853

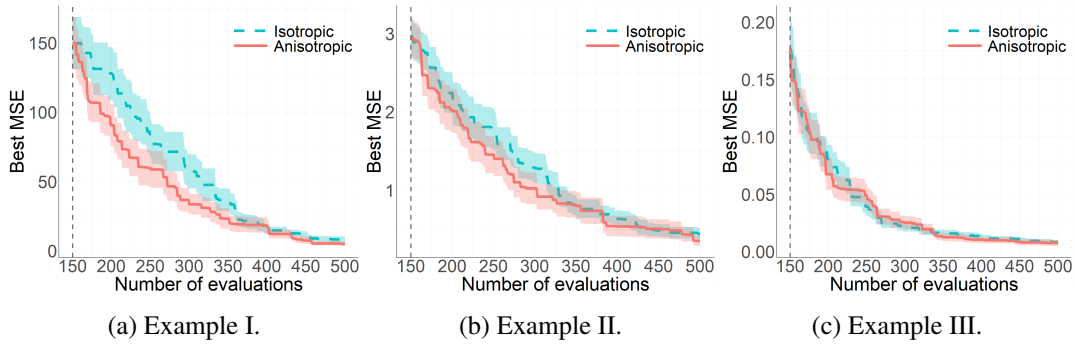


Fig 8: Trajectories of the lowest MSEs using isotropic- and anisotropic-kernel GPs in Ex-Sliced-Calib in the numerical study (Note: the band denotes the ± 1 standard error).

The deformation-based GP introduces nonstationarity by transforming the original input space Θ into a latent warped input space $U = g(\Theta)$, where the deformation function $g(\cdot)$ defines how spatial relationships are locally stretched, compressed, or rotated. A stationary GP is then placed on the transformed inputs, allowing the covariance structure in the original space to vary smoothly with location. This approach provides an intuitive geometric interpretation of nonstationarity, as correlations adapt according to the spatial deformation implied by $g(\cdot)$. Deformation-based GPs are particularly effective for capturing complex spatial dependencies and anisotropy in irregular domains. In our implementation, to apply a deformation-based GP, it first learns a spatial deformation using thin-plate splines (Sampson and Guttrop, 1992) that warps the inputs θ to u and then fits an isotropic GP on the warped inputs u .

Table 11 below summarizes the results, and Figure 9 below shows the corresponding trajectories of the best loss values over the rounds. Ex-Sliced-Calib using the stationary anisotropic GP consistently achieves lower test MSEs than the version using the deformation-based GP as the loss function surrogate. This is because the calibration loss surfaces in our study exhibit strong effect sparsity and are well characterized by axis-aligned anisotropic structure. In this setting, the anisotropic stationary kernel with ARD provides sufficient flexibility given the limited number of BO evaluations available.

In contrast, the deformation-based GP requires estimating substantially more degrees of freedom, such as warping maps, which cannot be reliably learned from the limited loss function evaluations, leading to unstable surrogates. In particular, in Example I, the deformation-based GP fails to accurately model the loss surface, causing Ex-Sliced-Calib to incorrectly

select the major subspace and misguide the subsequent design, thereby degrading calibration performance. These additional experiments support using the anisotropic stationary GP as a more reliable and accurate surrogate for our calibration setting compared to the nonstationary GP.

TABLE 11

Comparison of test MSEs and calibrated parameter values from 20 experiments using stationary and nonstationary GPs in Ex-Sliced-Calib in the numerical study (Note: the values inside parentheses are standard deviations, and the best results are highlighted in bold font).

Example	Kernel	Test MSE f	Calibrated Parameter Values		
			θ_1	θ_2	θ_3
I	Anisotropic	4.976 (7.405)	420.797 (5.542)	420.740 (3.284)	—
	Deformation	121.015 (84.304)	243.788 (288.429)	278.120 (296.146)	—
	<i>True parameter values</i>		420.969	420.969	—
II	Anisotropic	0.355 (0.295)	0.018 (0.064)	-0.013 (0.047)	—
	Deformation	0.604 (0.637)	-0.021 (0.067)	0.033 (0.108)	—
	<i>True parameter values</i>		0	0	—
III	Anisotropic	0.007 (0.012)	0.109 (0.019)	0.553 (0.005)	0.850 (0.008)
	Deformation	0.006 (0.005)	0.116 (0.012)	0.556 (0.008)	0.854 (0.006)
	<i>True parameter values</i>		0.115	0.556	0.853

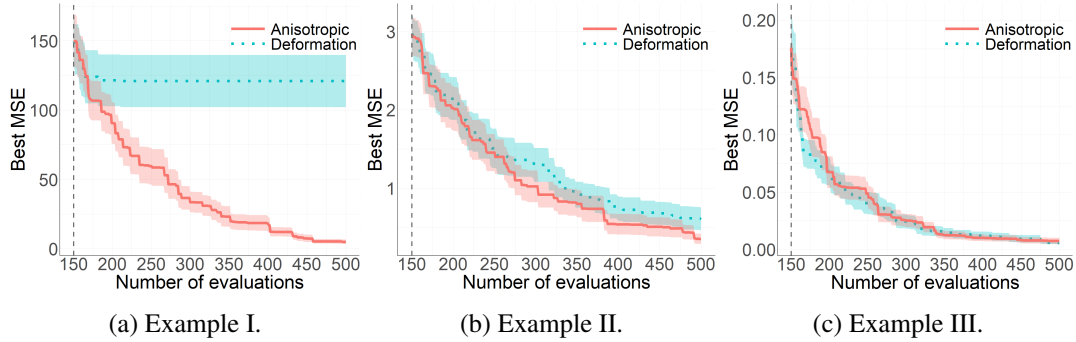


Fig 9: Trajectories of the lowest MSEs using Ex-Sliced-Calib with different kernel types (i.e., stationary vs. nonstationary) in the numerical study (Note: the band denotes the ± 1 standard error).

J. Numerical Study of Uncertainty Quantification for Estimated Parameters. To demonstrate how we quantify the uncertainty of the estimated parameters, we introduce a new example that simplifies the setting of the original Example 3 in the numerical study. This simplification is motivated by two main considerations. First, constructing CIs for the estimated parameters based on asymptotic normality requires the algorithm to reach the global minimum, i.e., $\theta_0^I = \hat{\theta}_{ML}^I$. However, the original numerical examples were deliberately designed as difficult optimization problems, making it challenging to consistently attain the global minimum under *limited* computational budgets. Second, these optimization problems are also computationally intensive to solve in each experiment. To reliably evaluate the empirical coverage rate, which measures the proportion of cases where the constructed CIs contain the true parameter values, a large number of experiments (e.g., 200) is required. In our original numerical study, we reported results from 20 experiments; scaling this up to 200 would be computationally prohibitive under the original settings.

For these reasons, we revisit Example 3 with modifications: we reduce the dimensionality from $D = 30$ to $D = 10$ and remove the $h(\cdot)$ component from the computer model, while keeping the number of true parameters d_I unchanged. As reported in Table 12 below, the test MSE in this simpler setting is nearly zero, and all parameter estimates are very close to their true values.

TABLE 12

Test MSE and calibrated parameter values from 200 experiments in Example III of the numerical study, obtained from Ex-Sliced-Calib with $D = 10$ and excluding $h(\cdot)$ (Note: the values inside parentheses are standard deviations).

Test MSE f	Calibrated Parameter Values		
	θ_1	θ_2	θ_3
0.0001 (0.0000)	0.115 (0.001)	0.556 (0.000)	0.853 (0.000)
<i>True parameter values</i>	0.115	0.556	0.853

Table 13 further summarizes the average half-length of the 95% CIs for each influential parameter and their empirical coverage rates from 200 experiments. Overall, the CIs frequently contain the true parameters, as the empirical coverage rates are close to the nominal 95% level. This demonstrates that our proposed method can provide high-quality uncertainty quantification results for the estimated parameters.

TABLE 13

Uncertainty quantification results from 200 experiments in Example III of the numerical study, obtained from Ex-Sliced-Calib with $D = 10$ and excluding $h(\cdot)$ (Note: the values inside parentheses are standard deviations).

Average Half Length of 95% CI			Coverage Rate (%)		
θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
0.012 (0.026)	0.001 (0.003)	0.001 (0.001)	96.0	97.0	94.5

We note that the computational challenges of our approach, particularly when solving high-dimensional problems with large D and incorporating the $h(\cdot)$ component, limit its applicability for uncertainty quantification analysis. This observation motivates the use of computationally cheaper surrogate models, rather than standard GPs, to further enhance the practicality of the proposed method. Such an extension is beyond the scope of this study and is left for future research.

K. Nugget Effect. When Ex-Sliced-Calib partitions the dimension into major and minor subspaces and constructs the GP surrogate model in the major parameter space, an issue may arise if distinct parameter sets in the original space are projected to the same point in the reduced space, potentially leading to different outputs. This phenomenon results in stochastic outputs within the reduced space.

To address this issue, we incorporate a nugget effect into the diagonal entries of the covariance matrix of the GP surrogate. The kernel function for each entry is defined as

$$(A.42) \quad k_{\text{nugget}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = k(\boldsymbol{\theta}, \boldsymbol{\theta}') + \tau \delta_{\boldsymbol{\theta}, \boldsymbol{\theta}'},$$

where $k(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denotes the base kernel function, such as the squared exponential or Matérn kernel, $\tau > 0$ is a nugget parameter, and $\delta_{\boldsymbol{\theta}, \boldsymbol{\theta}'}$ is the Kronecker delta. The nugget can either be set to a default value if the variance is known *a priori*, or treated as a hyperparameter to be estimated via maximum likelihood estimation (MLE) or cross-validation. This nugget term accounts not only for potential observational noise but also for the variability introduced by the non-injective nature of the projection, that is, cases where different parameter sets in the

original space are mapped to the same point in the reduced space but yield distinct outputs. In this study, we estimate τ in each round when constructing the GP in the reduced space using MLE.

This formulation allows the GP to capture a certain degree of local uncertainty at each input point in the reduced space, reflecting the inherent stochasticity in such cases. The use of a nugget is also known to improve numerical stability in GP inference and is widely adopted when modeling systems with latent or unmodeled sources of variation (Gramacy, 2020).

REFERENCES

- ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, U.S. Department of Commerce.
- FRAZIER, P., POWELL, W. and DAYANIK, S. (2009). The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing* **21** 599–613.
- GRAMACY, R. B. (2020). *Surrogate: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC.
- JONES, D. R., SCHONLAU, M. and WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13** 455–492.
- KUSHNER, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering* **86** 97–106.
- LYU, Y., YUAN, Y. and TSANG, I. W. (2019). Efficient batch black-box optimization with deterministic regret bounds. *arXiv preprint arXiv:1905.10041*.
- NEAL, R. M. (2012). *Bayesian learning for neural networks* **118**. Springer Science & Business Media.
- RASMUSSEN, C. E. and WILLIAMS, C. K. (2006). *Gaussian processes for machine learning*. MIT Press.
- SALTELLI, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications* **145** 280–297.
- SAMPSON, P. D. and GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association* **87** 108–119.
- SCHMIDT, A. M. and O’HAGAN, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **65** 743–758.
- SRINIVAS, N., KRAUSE, A., KAKADE, S. and SEEGER, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on Machine Learning* 1015–1022.