Robust and expert-agnostic digital twin calibration via ensemble learning and Bayesian optimization

Sicheng Zhan
szhan@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Bosen Cui
bosen_cui@mail.tsinghua.edu.cn
Tsinghua University
Beijing, China

Abstract

Digital twins have emerged as a critical tool in tackling climate change. Considering the data scarcity of complex systems, a promising approach to developing digital twins involves combining physics-based models with data assimilation. However, model calibration remains challenging due to uncertainties in both the physical models and observational data, and the reliance on domain knowledge. In this study, we develop an ensemble learning-based approach that aggregates sub-models with diversified calibration configurations. The proposed method streamlines calibration without expert-driven parameter screening and improves the digital twin's extrapolation capability, enabling more robust predictive applications. We demonstrate the effectiveness of our approach by calibrating the energy model of an office building, significantly reducing the extrapolation error and the associated risks. To the best of our knowledge, this is the first study to facilitate the calibration of physics-based models using ensemble learning, especially in the parameter space.

CCS Concepts

• **Theory of computation → Theory and algorithms for application domains**; • **Computing methodologies → Modeling and simulation**.

Keywords

model calibration, ensemble learning, Bayesian optimization, scientific machine learning, digital twin

ACM Reference Format:
Sicheng Zhan and Bosen Cui. 2025. Robust and expert-agnostic digital twin calibration via ensemble learning and Bayesian optimization. In *The 12th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BUILDSYS '25), November 19–21, 2025, Golden, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3736425.3770105

1 Introduction

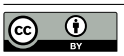
Digital twins are computational models that replicate the behavior of complex physical systems and support decision-making through virtual experiments. They have shown great potential in many applications that tackle climate change, including but not limited to optimizing the design and operations of energy systems considering the impact of global warming, analyzing the climate resilience of infrastructure systems, and projecting future climate change scenarios.

As the foundation of these applications, digital twins should be capable of 1) capturing the time-varying dynamics of the target systems, and 2) predicting unseen scenarios or unobserved variables in decision-critical regions [6]. Despite the progress in data acquisition technologies during the past decades, large and complex systems are still and will continue to be under-sampled, which is the bottleneck for purely data-driven models to succeed as digital twins. Therefore, a typical approach to satisfying the extrapolation requirements is to establish a physics-based model and learn the heterogeneous system dynamics from measured data [11].

The inverse problem of assimilating data and inferring the physical parameters provides a systematic approach to integrate the complementary strengths of real-world data and physics-based models. While many algorithms have been developed to calibrate digital twins and quantify the uncertainties [2–4], the procedure of defining and solving a calibration problem remains challenging and highly expert-driven. Uncertainties inevitably come from both the physics-based model and measured data, and it is intractable to simply calibrate all the parameters at the same time. As existing sensitivity analysis approaches suffer defects in accounting for the potential model discrepancy and parameter interactions, it is cumbersome to select the parameters and define the calibration task. Moreover, even with a reduced parameter space, the calibration is almost doomed to be ill-posed given the lack of high-resolution data to constrain the solution [12]. Although the identifiability issue may not stop the model from achieving satisfactory accuracy, the low predictive error does not guarantee the reliability of digital twins, causing problems in downstream applications [13].

Recent advances in ensemble learning and model aggregation have demonstrated improved model performance compared with single models for many applications [1, 7, 10]. Aggregating a group of diversified sub-models brings various benefits such as avoidance of overfitting, reduction of variance, and resistance to outliers. However, previous studies constructed sub-models mainly through purely data-driven methods, overlooking the great yet untapped potential of combining digital twins. Ensemble learning offers two significant opportunities for model calibration. First, subsets of uncertain parameters can be calibrated in different sub-models, saving the cost of screening while adding to the diversity. Further, the aggregation complements the governing equations of digital twins, yielding more robust extrapolation in decision-critical cases.

To the best of our knowledge, this is the first study that facilitates the calibration of physics-based models through parameter-space ensemble learning. In the following sections, we first illustrate the problem of traditional methods by calibrating the digital twin of an



BUILDSYS '25, Golden, CO, USA

ACM ISBN 979-8-4007-1945-5/2025/11
https://doi.org/10.1145/3736425.3770105

office building via Bayesian optimization (it can be other calibration methods). Then, we explain the technical details of ensemble learning integration and manifest its benefits using the same case study. Furthermore, a novel model aggregation method is proposed to account for the variance of sub-models and enhance the model interpretability. We conclude by summarizing the contributions and pointing out the directions for further development.

2 Traditional calibration via Bayesian optimization

For ease of illustration, we first define the inverse problem of digital twin calibration as Equation 1:

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

The true parameters $\boldsymbol{\theta}^*$ are sought by minimizing the objective function J. The first element $\hat{Y}$ denotes the target system responses predicted by the physics-based model $\mathcal{M}$, where $\boldsymbol{\theta}$ denotes the model parameters to be calibrated and X denotes the model inputs including disturbances and control actions. The second element Y denoted the observed ground truth of $\hat{Y}$, and the calibration is subject to the admissible parameter space Θ.

In search of $\boldsymbol{\theta}^*$, Bayesian optimization (BO) performs the exploration and exploitation trade-off with an acquisition function $\mathcal{A}$. The acquisition function uses the predictive distribution given by the surrogate model to compute the expected gain in minimizing J at each $\boldsymbol{\theta}$ sampled from the parameter space. The next $\boldsymbol{\theta}$ to evaluate is given by maximizing $\mathcal{A}$.

Figure 1 displays the results of a simple calibration experiment using the DOE prototype building models[1], where a baseline model was calibrated to capture the behavior of a similar but more energy-efficient building. Emulating the data availability in an actual building, six parameters[2] were selected for calibration, and the target model output was hourly electricity consumption. The left plot shows that BO quickly converged after the initialization stage. Correspondingly, the electricity was accurately predicted during the testing period[3]. However, the calibrated parameters $\boldsymbol{\theta}^*$ were deviated from the true values, mapped to the drifted indoor temperature trajectory.

3 Robust calibration via ensemble learning

Figure 2 illustrates the workflow of calibrating a physics-based digital twin via Bayesian optimization and ensemble learning. We first establish a group of sub-models through parameter sampling, which are then fed to the BO-based calibration, leading to a unique set of calibrated parameters for each sub-model. When drawing predictions, the outputs of calibrated sub-models are aggregated

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

[2]The calibrated parameters included envelope thermal resistances, outdoor airflow rates, infiltration rates, equipment power densities, solar heat gain coefficients, and cooling setpoints.

[3]The prediction is almost perfect because the target building is also synthetic, making the discrepancy in model structure minimal and the uncalibrated parameters similar to the target. The post-calibration discrepancy will be more significant in a more uncertain real-world case, but the demonstrated issues will persist.

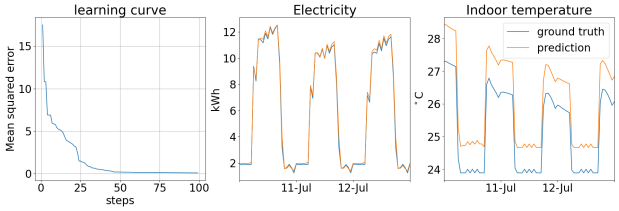


Figure 1: Results of a single digital twin calibration using Bayesian optimization.

through Exponentially Weighted Averaging (EWA)[8]. In this section, we elaborate on the process of sub-model development and post-calibration aggregation, and present the improved model performance when calibrating the previous prototype energy model.

In Equation 1, $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the number of parameters to calibrate. For each sub-model in ensemble learning, we sample a subset of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_S \in \mathbb{R}^{|S|}$ for calibration, $|S| \leq d$, and fix the other parameters at their initial values. Thereby, the j^{th} sub-model can be denoted by $\mathcal{M}(X, \boldsymbol{\theta}_{S_j})$. Depending on the number of uncertain parameters and the computational resources, random sampling, combinatorial enumeration, or other sampling approaches could be adopted to determine $\boldsymbol{\theta}_{S_j}$. Yet, there are two general principles: each subset should include more than two parameters to account for the interaction and ensure the expressivity of sub-models ($|S_j| > 2$), and the union of all the k subsets should ideally include all the candidate parameters to avoid the loss of information ($|\bigcup_{j=1}^{k} S_j| = d$).

After the calibration of each sub-model, the weighted average of all sub-model outputs gives the ensemble predictions. A common way to determine the weights is to average all sub-models ($\pi_j = 1/k$). Although this approach outperformed a single model in many empirical studies, it relies on fairly good sub-models. In the streamlined calibration without prior domain knowledge, a sub-model could be incapable of fitting the data, resulting in bad predictions and meaningless parameter values. In such cases, the averaging weights should be modified to diminish its impact on the final prediction. Rationally, it is non-optimal to assign equal weights to an underfitted model and a well-trained model. On the other hand, even with relatively efficient methods such as Bayesian optimization, it is expensive to find the globally optimal weights to aggregate the models.

To this end, we propose an adaptive strategy, namely Exponentially Weighted Averaging (EWA), to assign near-optimal weights without solving another complicated optimization. EWA assigns weights for the k sub-models according to the following equation:

$$\begin{aligned} \hat{\omega}_j &= \frac{\pi_j e^{-\beta \hat{r}_j}}{\hat{\mathcal{L}}}, \qquad j = 1, \cdots, k \\ \hat{\mathcal{L}} &= \sum_{j=1}^{k} \pi_j e^{-\beta \hat{r}_j}, \end{aligned} \tag{2}$$

where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_k)^T$, is a known distribution on the weight space, $\beta > 0$ controls the degree of concentration of the model

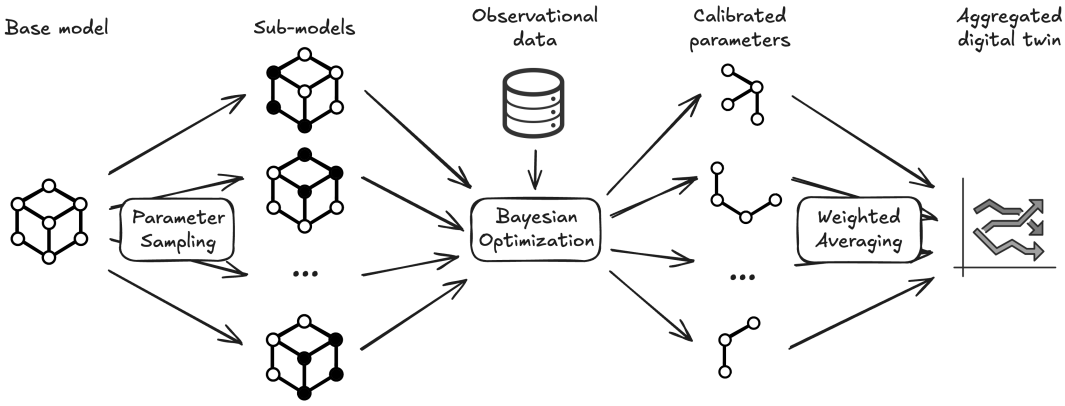


Figure 2: Workflow of digital twin calibration via Bayesian optimization and ensemble learning.

weights, $\hat{r}_j$ is an unbiased estimate of the risk of the j^{th} sub-model, which can be approximated by mean squared error (MSE).

EWA assigns model weights based on the smoothed version of the negative estimated risk, which outperforms direct use of the negative estimated risk in high-dimensional settings. As a result, larger weights are given to more representative sub-models, while those with higher risks can still contribute to the aggregation. $\boldsymbol{\pi}$ can be viewed as a prior probability in the model space where each π_j reflects how much prior belief we have about the jth model. When this prior knowledge is unavailable, a uniform distribution can be adopted. It is worth mentioning that the tuning parameter β is crucial. When $\beta \to 0$, the EWA weights will essentially converge to $\boldsymbol{\pi}$ the known prior distribution. On the other hand, the weights can lead to a uniform distribution of the weight space.

3.1 Experimental setup and results

Using the same base model as the previous example, sub-models were created by only calibrating four of the original six parameters. Covering all the combinations of four parameters yielded 15 sub-models. For consistency, BO-based calibration was applied for each sub-model. As shown in the left plot in Figure 3, while some cases converged faster, the learning curves of most sub-models were close to the single calibration with all six parameters. As expected for ensemble models, the total computational time was much longer (around ten times in this case). However, considering the curse of dimensionality, the extra computational cost may shrink as the total number of parameters increases.

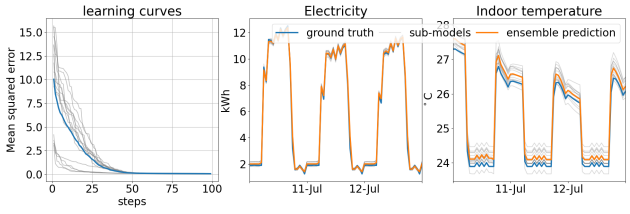


Figure 3: Results of digital twin calibrations via Bayesian optimization and ensemble learning.

Similar to single calibration, each sub-model predicted the target output almost perfectly during the testing period. This manifests the better extrapolation capability of physics-based models. Although the parameters are likely unidentifiable, the governing equations still constrain the optimization to capture the physical relationships between the disturbances and the target outputs. By contrast, the extrapolative prediction of indoor temperature (average of five rooms) varied among the sub-models, reflecting the intricate interaction between the parameters and the physics-based model when learning from the data. This also highlights the risk of justifying a calibrated model only based on the accuracy of target output prediction. For example, the potential to reduce cooling energy could be underestimated given a high indoor temperature.

As expected, some sub-models overestimated the indoor temperature, while some others underestimated it. Aggregating them canceled out the errors and resulted in the accurate extrapolation as shown in the right plot. Some randomly sampled parameter combinations yielded noticeably accurate extrapolations, but it is worth noting that we cannot simply use these models because their merit was only uncovered with the ground truth data. Consequently, the MSE of temperature prediction was significantly decreased from 0.74°C (single calibration) to 0.05°C.

4 Sub-model development and interpretation

Table 1 summarizes the key differences between traditional ensemble learning based on purely data-driven models, traditional single physics-based model calibration, and the proposed ensemble learned physics-based model calibration. The proposed method takes advantage of the two traditional methods in terms of predictive capability, expert dependency, and interpretability. It has been acknowledged that ensemble methods based on black-box models have a strong capability of fitting the data but unreliable for extrapolation. As for single physics-based model calibration, it is generally more restricted with a lower degree of freedom in training but is known to be more extrapolatable and interpretable. Yet, they highly rely on experts to configure the calibrations and characterize the model-to-reality discrepancy. As demonstrated in Figure 1, an ill-posed calibration task could result in a physically drifted model with deceptively low predictive errors.

Table 1: Key difference between traditional ensemble learning based on purely data-driven models, traditional single physics-based model calibration, and the proposed ensemble learned physics-based model calibration.

	Black-box ensemble learning	Single physics-based model calibration	Ensemble learned physics-based model calibration
Fitting capability	High	Medium	High
Extrapolation capability	Low	Medium	High
Expert dependency	Low	High	Low
Interpretability	Low	High	High

In contrast, the proposed method provides satisfactory extrapolation without extensive expert interference. Compared with traditional ensemble learning methods, the aggregation of physics-based sub-models first preserves the flexibility in training and carries over the capability of predicting unseen variables. Furthermore, calibrating a group of models relaxes the requirements of expert knowledge to define each optimization problem. The various parameter combinations when calibrating sub-models serve as different constraints in the optimization problems[4], improving identifiability and fitting the training data with different system dynamics. Some combinations could result in models that perform worse, which is reflected in the EWA weighting factors and can be iterated as priors to enhance model reliability. This also adds to the interpretability by providing information about the importance of alternative variables or sub-models, as well as the interrelationships.

Developing sub-models in the parameter space amplifies the synergy between ensemble learning and physics-based models. The traditional approaches to developing sub-models for ensemble learning can be categorized into homogeneous and heterogeneous. Homogeneous ensemble learning uses the same type of model repetitively but with different training configurations, such as by sampling different subsets of training data. For example, Liu et al. [5] also investigated the application of ensemble learning for model calibration, but only by varying training periods. On the other hand, heterogeneous ensemble learning combines different types of models, leveraging their unique strengths and mitigating their weaknesses [9]. The diverse sub-models are expected to complement each other and capture a wider variety of data patterns. The proposed method adds a new dimension to the heterogeneity, where different combinations of parameters respectively vary different parts of the complicated dynamics.

Beyond the proof-of-concept example, it is unnecessary to include all the possible combinations in a real-world case with a larger number of candidate parameters, nor to have the same number of parameters in each sub-model. The proposed method is also compatible with other sub-modeling methods. For example, sub-models with different parameter sets can be calibrated with different training data. Instead of random resampling in bootstrapping, the dataset can also be strategically resampled based on seasons or other strategies. Furthermore, although expert-agnostic, domain knowledge can still be incorporated by training sub-models using different combinations of input-output pairs that are more relevant to the selected parameters. This could exclude irrelevant models beforehand, thereby reducing computational costs.

[4]Technically, single model calibration can also be constrained for better identifiability, but the information, such as a narrow range of parameter value, is typically unavailable.

5 Impact and future work

Tackling climate change involves many open and complex systems, such as infrastructures, grids, and the climate system itself, requiring digital twins with reliable extrapolation for robust decision-making. We present a robust calibration method using ensemble learning, which avoids expert-driven parameter screening and enhances the extrapolation capability of calibrated digital twins.

The proposed method is demonstrated through an office building energy model, paving the way for broader applications in complex and data-scarce open systems. Further investigation mainly lies in the development of sub-models. While we currently focus on physics model parameters $\boldsymbol{\theta}$, there is an opportunity to optimize the variations of X and Y to maximize the information gain of each sub-model and improve the computational efficiency.

References

[1] Seyed Babak Haji Seyed Asadollah, Ahmad Sharafati, and Shamsuddin Shahid. 2022. Application of ensemble machine learning model in downscaling and projecting climate variables over different climate regions in Iran. *Environmental Science and Pollution Research* 29 (2022), 1–20.

[2] Ankush Chakrabarty, Emilio Maddalena, Hongtao Qiao, and Christopher Laughman. 2021. Scalable Bayesian optimization for model calibration: Case study on coupled building and HVAC dynamics. *Energy and Buildings* 253 (2021), 111460.

[3] Dagimawi D Eneyew, Miriam AM Capretz, and Girma T Bitsuamlak. 2024. Continuous model calibration framework for smart-building digital twin: A generative model-based approach. *Applied Energy* 375 (2024), 124080.

[4] Marc C Kennedy and Anthony O'Hagan. 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 3 (2001), 425–464.

[5] Y Liu, Z Liu, S Zhang, R Jacob, F Lu, X Rong, and S Wu. 2014. Ensemble-based parameter estimation in a coupled general circulation model. *Journal of Climate* 27, 18 (2014), 7151–7162.

[6] Steven A Niederer, Michael S Sacks, Mark Girolami, and Karen Willcox. 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science* 1, 5 (2021), 313–320.

[7] Vahid Nourani, Ali Hasanpour Ghareh Tapeh, Kasra Khodkar, and Jinhui Jeanne Huang. 2023. Assessing long-term climate change impact on spatiotemporal changes of groundwater level using autoregressive-based and ensemble machine learning models. *Journal of Environmental Management* 336 (2023), 117653.

[8] Philippe Rigollet and Alexandre B Tsybakov. 2012. Sparse Estimation by Exponential Weighting. *Statist. Sci.* 27, 4 (2012), 558–575.

[9] Zeyu Wang, Zhixi Liang, Ruochen Zeng, Hongping Yuan, and Ravi S Srinivasan. 2023. Identifying the optimal heterogeneous ensemble learning model for building energy prediction using the exhaustive search method. *Energy and Buildings* 281 (2023), 112763.

[10] Zeyu Wang, Yueren Wang, and Ravi S Srinivasan. 2018. A novel ensemble learning approach to support building energy use prediction. *Energy and Buildings* 159 (2018), 109–122.

[11] Karen E Willcox, Omar Ghattas, and Patrick Heimbach. 2021. The imperative of physics-based modeling and inverse theory in computational science. *Nature Computational Science* 1, 3 (2021), 166–168.

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

[13] Sicheng Zhan, Mingya Zhu, Siyu Cheng, and Adrian Chong. 2024. Bridging performance gap for existing buildings: The role of calibration and the cascading effect. *Building Simulation* 18, 1 (2024), 123–140.