OPEN ACCESS

LETTER

Journal homepage: www.the-innovation.org/energy-use

Energy Use

INNOVATION PRESS

Energy Use 1 (2025):100008

Contents lists available at INNOVATION PRESS

Generalizable Data-Driven Building Operation Control Methods for Scalable Cross-Building Applications

Cheng Fan,[1,2] Yutian Lei,[2] Marco Savino Piscitelli,[3] and Jinhan Mo[1,2,*]

[1]State Key Laboratory of Subtropical Building and Urban Science, Shenzhen University, Shenzhen 518060, China

[2]Key Laboratory of Coastal Urban Resilient Infrastructures, College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China

[3]Department of Energy, TEBE research group, BAEDA Lab, Politecnico di Torino, Italy

*Correspondence: mojinhan@szu.edu.cn

Received: August 22, 2025; Accepted: September 5, 2025; Published Online: September 25, 2025; https://doi.org/10.59717/ipj.energy-use.2025.100008

Citation: Fan C., Lei Y., Piscitelli M. S., et al. (2025). Generalizable Data-Driven Building Operation Control Methods for Scalable Cross-Building Applications. *Energy Use* **1**:100008.

Dear Editor,

Building operations account for approximately 30% of global final energy use[1] and continue to exhibit an upward trend. To facilitate the goal of carbon neutrality in the building sector, it is urgently needed to develop intelligent and scalable building operation technologies to better regulate building energy efficiency.

The operation of modern buildings emcompasses the production, distribution, and consumption processes of multiple energy flows, such as electricity, gas and thermal energy, and has been gradually evolving into a highly integrated and complex system. Conventional building operation methods primarily rely on engineering experience and human labor, making it difficult to derive optimal operation settings of numerous control nodes in constantly changing operating conditions. With the advancement in information technologies and artificial intelligence (AI), data-driven methods have gained great interest to upgrade conventional human-centric operation approaches in an automated fashion and thereby, enhancing the efficacy of various building operation tasks, such as anomaly identification, energy load forecasting, component- and system-level optimal control.

However, the present application of data-driven building operation technologies is rather limited, especially when deployed to building systems with heterogeneous architectures and configurations. This letter aims to analyze the key challenges in developing generalizable methods for scalable cross-building applications, while proposing possible data-driven solutions to tackle such challenges.

KEY CHALLENGES IN THE GENERALIZABLE AND SCALABLE DEPOLYMENT OF INTELLIGENT BUILDING OPERATION TECHNOLOGIES

At present, data-driven building energy management methods have yet to be widely applied, primarily due to the following three common challenges.

Limited human resources resulting in unreliable data ground truths

The fundamental basis for developing data-driven building operation technologies is to differentiate normal and abnormal operations in historical data. While data representing normal operations can be used to train models to simulate expected building system behaviors given varying operating conditions, the discovery of abnormal operation data can provide clues for amending faulty or suboptimal operations. In practice, only limited human resources are typically available to conduct such data labeling, leaving the majority of building operational data unlabeled for the direct implementation of learning algorithms.[2]

Inconsistent building data quality and semantics leading to inaccurate data-driven models

Variations in building functionalities and operating conditions often result in inconsistent data quality with varying monitoring variables. The lack of sematic sufficiency, i.e., inconsistent descriptions of variables, units and contextual metadata, further reduces the interoperability of building datasets. These issues compromise the precision of data-driven models and inevitably introduce decision biases in the control strategies built upon them. For example, chiller operating efficiency is closely related to their part-load ratios. Given accurate building cooling load predictions, the running chiller numbers can be optimized to maximize chiller operating efficiencies, while inaccurate predictions will lead to rather poor chiller sequencing controls.[3]

Heterogeneous building systems requiring customized control strategies

Owing to differences in building types, environmental conditions and indoor demands, different buildings may adopt energy systems with various architectures and configurations, making it difficult to develop generalizable and efficient optimal control strategies. Taking buildings equipped with photovoltaic (PV) and energy storage systems as an example, the temporal relationships among energy consumption, PV generation, and storage capacities are constantly changing. It can be highly challenging to define universal control rules for energy storage systems to achieve optimal controls given varying time-of-use tariffs and operating conditions.

OUTLOOK

Extensive research has been conducted to show the power of statistics, machine learning and data mining techniques in extracting, representing and adjusting building operation patterns for enhanced energy efficency.[4] Despite encouraging progress, most studies are tailored to specific building systems and rely on the assumption that high-quality data is available, limiting their scalability for cross-building applications.

From the authors' perspectives, generalizable data-driven building operation technologies should have three desired properties: (1) requires minimized human labor to differentiate between normal and abnormal building operation data; (2) can provide reliable predictions on building system performance given limited data at individual building level; (3) has the ability to provide efficient and stable optimizations for complex and heterogenous systems. Based on the above-mentioned, we propose a three-stage analytic framework, which leverages multiple machine learning paradigms, to facilitate data-driven building operations. As shown in Figure 1, the framework

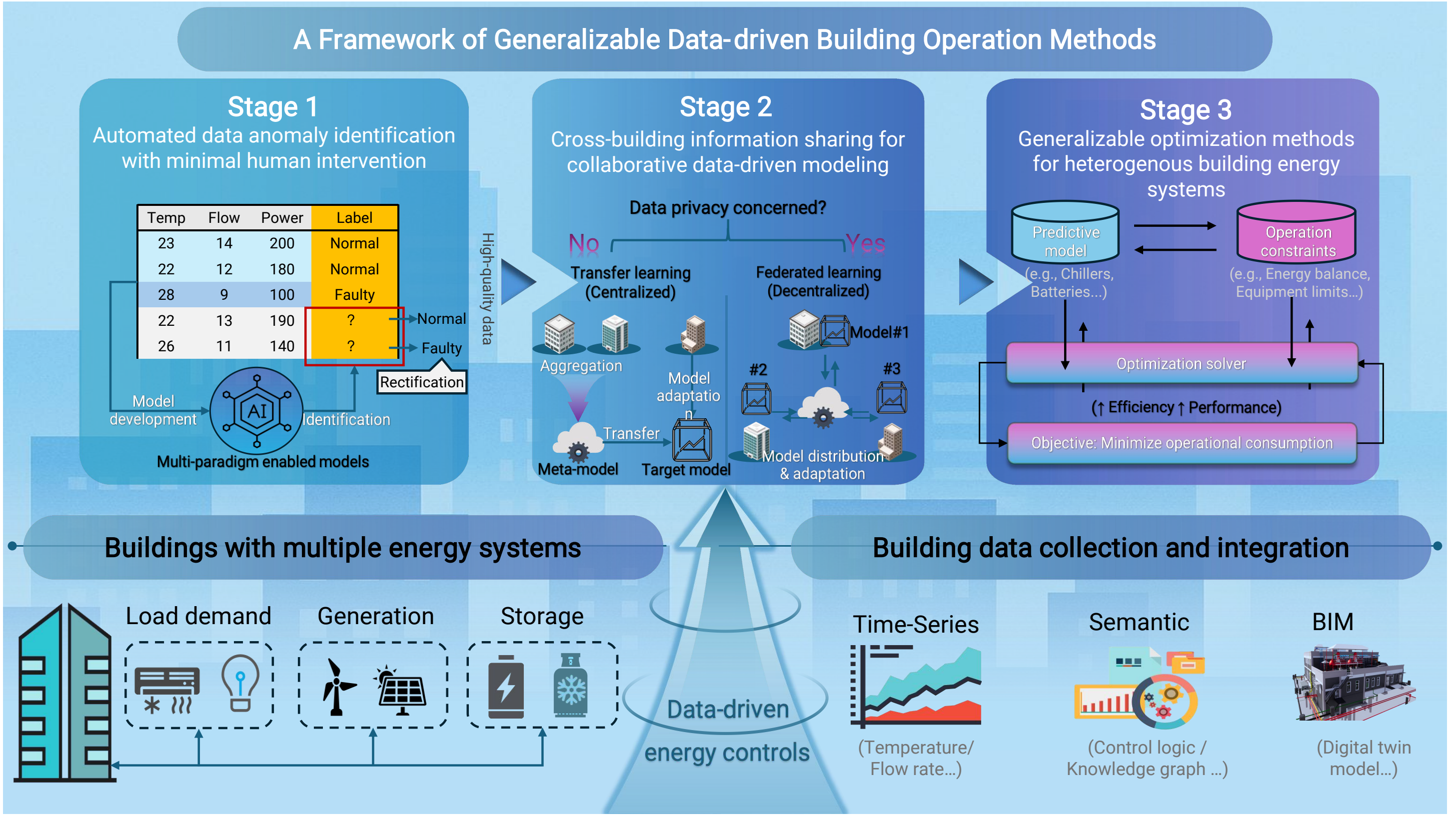

Figure 1. A three-stage analytic framework for generalizable data-driven building operation methods The framework integrates multiple machine learning paradigms to achieve anomaly identification, predictive modeling and control optimization with minimal human labor and data requirements

establishes a comprehensive workflow spanning the entire decision-making pipeline with three stages, i.e., data anomaly identification, predictive model development, and building system control optimization, creating a synergistic data-driven mechanism to minimize human labor and data requirement for applications.

Multiple machine learning paradigms-enable methods for data anomaly identification with reduced human labor

Due to the wide existence of sensor malfunctions, equipment faults and infrequent human activities, building operation data collected typically consist of both normal and abnormal data. The first stage focuses on developing an efficient data anomaly identification system with reduced human labor. Taking the classification-based approach for anomaly detection as an example, conventional supervised learning methods typically require extensive labeled data for model training. To reduce human labeling costs, multiple machine learning paradigms can be integrated to facilitate essential steps in developing anomaly classification models. For instance, active learning creates an efficient iterative process between data-driven models and experts to automatically select informative data samples for manual inspection, thereby reducing the human labor required for data labeling. Informative data samples are those that may create the largest confusion (e.g., measured by least confidence or entropy) to the existing data-driven model and including them in the model training process will likely result in the largest improvement in model update.

As another example, considering the imbalanced nature of normal and abnormal operations, specific measures are typically needed to enhance the reliability of data-driven anomaly identification. Besides conventional data resampling and cost-sensitive methods, generative learning models, such as generative adversarial networks or variational autoencoders, can be developed to generate synthetic yet physically meaningful minority-class data to enhance the reliability of data-driven models. Other machine learning paradigms, such as self-supervised and semi-supervised learning, can integrate both labeled and unlabeled data for anomaly identification and thereby reduce the dependency on human labor for decision-making. The combined use of these paradigms could achieve synergistic effects. For instance, previous studies have demonstrated that a hybrid approach integrating active learning and semi-supervised learning can reduce human labeling costs by 50%.[5]

Centralized and decentralized information sharing mechanisms for accurate data-driven modeling with limited individual building data

The generalizability of data-driven models is highly sensitive to the representativeness of training data, e.g., a model trained by summer operation data tends to fail in winter conditions. As a result, data-driven approaches may not be feasible for different buildings, as their operation data may vary in coverage due to limited operation time or operating conditions. Therefore, the second stage aims to break the data silo in the building field and utilize operational data from existing buildings for collaborative model development.

Existing studies have explored the potential of both centralized and decentralized methods for cross-building information sharing.[6] The centralized method can be formulated using the idea of transfer learning, and it is particularly suitable when building operation data can be exchanged without privacy concerns. The general idea is to pre-train a model using data-rich source buildings and then fine-tune the model using limited target building data.[7] The efficacy of such an approach has been validated in prior studies. For instance, in 24-hour ahead building energy prediction tasks across five distinct building types, the transfer learning-based approach can reduce prediction errors by more than 15%, compared with models trained on individual data alone.[8] Meanwhile, in scenarios where data privacy regulations prohibit data centralization, federated learning offers a privacy-preserving alternative. Such an information-sharing mechanism is especially valuable for cross-stakeholder collaborations, enabling collective model training without data exchange. In federated learning, each building trains a model locally and only transmits model updates (e.g., model gradients or weights) to a central server for global model aggregation, which will be re-distributed to all individual buildings for next-round training. Previous studies have shown that federated learning could improve building energy prediction accuracies by more than 5.0%.[9] While the performance enhancement of federated learning-based methods may not match their centralized counterparts, it does outperform conventional localized methods while addressing data privacy concerns.

Stable and efficient optimization methods for heterogenous and complex building energy systems

Ideally, given accurate data-driven prediction models on building energy systems, optimization methods can be developed to find optimal settings of numerous control nodes for enhanced energy efficiency. In practice, the key challenge at this stage is to develop stable and efficient optimization methods for complex and heterogenous systems. Firstly, less randomness in optimization results is always desired, as frequent variations in control settings may not be feasible or will bring negative impacts on equipment life and safety. For example, under the same environmental conditions and predicted building cooling loads, the optimization should yield consistent chiller and cooling tower operating schedules across consecurtive time steps to avoid frequent one-off switching. In such a case, algorithms that are vulnerable to local optima may not be suitable for practical applications.

Secondly, as data-driven optimizations may introduce complicated interactions among different data-driven models, the computation burden can be rather high for real-time implementations. In this case, domain expertise is vital to enhance the optimization efficiency. Taking the optimal control of a building system equipped with both PV and battery storage systems as an example, a model predictive control framework can be formulated to optimize the battery actions for minimized operation costs given time-of-use tariffs, and predictions on building energy demands and PV generations. It can be solved as a fix-horizon linear programming problem, yet the number of decision variables and constraints can be rather high in conventional settings, e.g., the decision variables are 96, considering a 24-hour prediction horizon and 15-minute time interval. Experienced building staff may realize that the battery charging or discharging status should be maintained unless either the direction of net energy (i.e., building energy consumption minus PV generation) or time-of-use tariff changes. In such a case, a simplified optimization task can be defined by aggregating temporally adjacent predictions according to their operating conditions, leading to around 90% reductions in the numbers of decision variables and constraints.[10]

CONCLUDING REMARKS: TOWARDS A WIDER IMPLEMENTATION OF DATA-DRIVEN SOLUTIONS

From our perspectives, the limited penetration of data-driven or AI-based solutions for smart building operations is no longer an algorithm-level issue. Instead, the main barriers lie in the lack of portability, interoperability, transferability, and also interpretability of these solutions across diverse building contexts and stakeholders. Equally important, the high costs associated with implementation and long-term maintenance create additional barriers to large-scale applications. Integration into existing building automation systems often requires bespoke engineering efforts, and demands ongoing monitoring, updates, and technical expertise. All these issues could discourage stakeholders from moving beyond pilot projects. Addressing these barriers requires a shift of focus: from developing ever more complex algorithms for specific tasks to creating standardized and cost-effective decision-making pipelines. Only through such efforts will data-driven or AI-based building operation technologies become trustworthy, transferable, and broadly impactful in practice.

REFERENCES

1. Gonzalez-Torres M., Perez-Lombard L., Coronel J.F., et al. (2022). A review on buildings energy information: Trends, end-uses, fuels and drivers. *Energy Rep.* **8**:626–637. DOI:10.1016/j.egyr.2021.11.280
2. Blum D., Candanedo J., Chen Z. et al. (2023). Data-driven smart buildings: State-of-the-art review. Candanedo J. (Ed.). *Energy in Buildings and CommunitiesTechnology Collaboration Programme* (CSIRO), pp:2537-2544. ISBN:978-1-4863-1894-0.
3. Sala-Cardoso E., Delgado-Prieto M., Kampouropoulos K., et al. (2020). Predictive chiller operation: A data-driven loading and scheduling approach. *Energy Build.* **208**:109639. DOI:10.1016/j.enbuild.2019.109639
4. Kathirgamanathan A., De Rosa M., Mangina E., et al. (2021). Data-driven predictive control for unlocking building energy flexibility: A review. *Renew. Sustain. Energy Rev.* **135**:110120. DOI:10.1016/j.rser.2020.110120
5. Fan C., Wu Q., Zhao Y., et al. (2023). Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance. *Appl. Energy* **356**:122356. DOI:10.1016/j.apenergy.2023.122356
6. Pinto, G., Piscitelli, M.S., Vazquez-Canteli, et al. (2021). Coordinated energy management for a cluster of buildings through deep reinforcement learning. *Energy* **229**:120725. DOI:10.1016/j.energy.2021.120725
7. Fan C., Sun Y., Xiao F., et al. (2020). Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Appl. Energy* **262**:114499. DOI:10.1016/j.apenergy.2020.114499
8. Pinto G., Messina R., Li H., et al. (2022). Sharing is caring: An extensive analysis of parameter-based transfer learning for the prediction of building thermal dynamics. *Energy Build.* **276**:112530. DOI:10.1016/j.enbuild.2022.112530
9. Fan C., Chen R., Mo J., et al. (2024). Personalized federated learning for cross-building energy knowledge sharing: Cost-effective strategies and model architectures. *Appl. Energy* **362**:123016. DOI:10.1016/j.apenergy.2024.123016
10. Fan C., Lu M., Sun Y., et al. (2024). A novel linear programming-based predictive control method for building battery operations with reduced cost and enhanced computational efficiency. *Renew. Energy* **237**:121847. DOI:10.1016/j.renene.2024.121847

FUNDING AND ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of this research by the National Natural Science Foundation of China (No. 52578137, 52278117), Guangdong Basic and Applied Basic Research Foundation (2024A1515011549) and Shenzhen Science and Technology Program (No. 20220531101800001, 20240813143330039).

DECLARATION OF INTERESTS

The authors declare no competing interests.