Proceedings of the 4th International Conference on Functional Materials and Civil Engineering
DOI: 10.54254/2755-2721/2026.34152

A Review of Data-Driven Methods for Smart Building Thermal Environments and HVAC Energy Optimization

Jiacheng Ding

School of Civil Engineering and Architecture, Zhejiang Sci-Tech University, Hangzhou, China
1197582047@qq.com

Abstract. Building operation accounts for a substantial share of global energy use and energy-related carbon emissions. Heating, ventilation and air-conditioning (HVAC) systems are central to this problem because they determine both indoor thermal conditions and operational energy demand. With the development of sensors, smart meters, building automation systems and energy management platforms, data-driven methods have become important tools for building load prediction, thermal comfort assessment, occupant behavior modeling and HVAC control optimization. This paper reviews data-driven methods for smart building thermal-environment regulation and HVAC energy optimization from the perspective of coordinated comfort and energy efficiency. The review first discusses how multi-source operational data describe weather disturbances, indoor environmental states, equipment responses, energy use and occupant demands. It then summarizes the roles of machine learning and deep learning in energy prediction, thermal-environment prediction, personal thermal comfort modeling and occupancy-related modeling. The paper further examines how prediction models are embedded in model predictive control, demand-controlled ventilation and reinforcement learning frameworks. The analysis shows that future progress should move beyond offline prediction accuracy and place greater emphasis on closed-loop performance, physical consistency, model transferability, occupant acceptance and practical deployment. A closer integration of building physics, operational data and occupant-centered control is essential for reliable low-carbon operation of smart buildings.

Keywords: smart buildings, data-driven methods, thermal environment, building energy consumption, HVAC, optimal control

1. Introduction

Energy consumption during building operation is closely related to indoor environmental regulation. The International Energy Agency reported that building operation accounts for approximately 30% of global final energy consumption and around 26% of global energy-related carbon emissions [1]. The GlobalABC report further indicates that the buildings and construction sector remains a critical domain for controlling global energy demand and carbon emissions [2]. Within building operation, HVAC systems provide heating, cooling, ventilation and dehumidification. Their operation strategies directly influence thermal comfort, indoor air quality and energy use. As urban density increases and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34152

expectations for indoor environmental quality rise, HVAC systems face more variable and more demanding operating conditions. Energy conservation in buildings therefore requires operational strategies that match energy supply with actual environmental and occupancy demand.

Traditional building energy analysis relies mainly on white-box models, empirical parameters and rule-based control. White-box models can represent heat and moisture transfer processes and are useful for design-stage performance assessment. During operation, however, building states are affected by weather fluctuations, envelope thermal inertia, equipment degradation, occupant activities and maintenance practices. Continuous model calibration is therefore difficult. Rule-based control and PID control remain widely used because they are reliable and easy to maintain, but they make limited use of future load variations, occupancy changes and electricity-price signals. Balali et al. [3] reviewed data-driven and hybrid models for building heating and cooling systems and emphasized that model accuracy, control integration and generalization remain key barriers. Li et al. [4] also showed that HVAC energy prediction research is shifting from algorithm comparison toward data quality, prediction scale, feature construction and operational adaptability.

Smart buildings provide a richer data basis for operational optimization. Sensors, smart meters, building automation systems and energy management platforms can continuously record indoor and outdoor conditions, equipment states, energy use and occupant activities. Data-driven methods can learn relationships among building thermal response, equipment regulation and occupant behavior from these measured data. These relationships support load prediction, indoor environment assessment and control optimization. Occupant-related factors make the problem more complex. Occupancy status, activity level, window or shading operation and thermal comfort preferences reshape indoor load evolution and control objectives [5, 6]. Thus, data-driven methods should be examined not only as prediction tools but also as components of operational decision-making.

This paper reviews data-driven methods for smart building thermal environments and HVAC energy optimization. The discussion is organized around three questions. First, how can multi-source operational data support joint prediction of load, indoor environment, air quality and occupant demand? Second, how can prediction models be embedded in HVAC control frameworks such as model predictive control, demand-controlled ventilation and reinforcement learning? Third, what limits the transition from offline prediction models to deployable control systems in real buildings? The objective is to clarify the functional roles, applicable boundaries and future development of data-driven methods in comfort and energy-efficiency coordination.

2. Operational data and thermal-environment energy prediction in smart buildings

Data-driven research in smart buildings is built on continuous observation of building operation. An operating building is a dynamic system influenced by weather disturbance, envelope heat storage, equipment regulation, occupant activities and maintenance strategies. External disturbances are transmitted to indoor spaces through the envelope and mechanical systems. Equipment actions change thermal and moisture states, while occupant activities reshape load demand. These factors jointly determine energy consumption, indoor air quality and comfort. Data-driven prediction should therefore provide information for feedforward regulation, constrained control and operational optimization rather than remain limited to statistical fitting.

2.1. Multi-source operational data and modeling value

Operational data in smart buildings are commonly obtained from weather stations, indoor sensors, building automation systems, energy metering platforms and occupant-sensing devices. Each source

DOI: 10.54254/2755-2721/2026.34152

describes a different part of the operational process. Weather variables represent external thermal disturbances. Indoor environmental variables describe space responses. Equipment variables record HVAC control actions. Energy variables reflect operating cost and energy intensity. Occupant variables connect internal load generation with comfort demand. A model trained only on energy series often captures statistical correlation rather than controllable system response. When equipment actions, indoor states and occupancy information are modeled together, prediction becomes more suitable for control-oriented representation.

Table 1. Multi-source operational data, modeling tasks and control value in smart buildings

Data category	Typical variables	Modeling tasks	Control value
Outdoor weather data	Outdoor temperature, humidity, solar radiation, wind speed	Cooling/heating load prediction and disturbance identification	Support pre-cooling, pre-heating and feedforward load adjustment
Indoor environmental data	Temperature, humidity, CO_2, air velocity, mean radiant temperature	Thermal-environment prediction, IAQ assessment and comfort evaluation	Constrain indoor environmental deviation and maintain air quality
HVAC equipment data	Fan frequency, damper opening, chiller power, supply/return air temperature	Equipment-state identification and system-response modeling	Optimize operating conditions and reduce inefficient operation
Energy and cost data	Total electricity use, sub-metered energy use, heating/cooling energy and electricity price	Energy prediction, anomaly detection and demand response	Evaluate energy-saving effects, operating cost and peak-shaving potential
Occupant data	Occupancy count, arrival/departure time, activity level and feedback	Occupancy prediction, behavior recognition and personalized comfort modeling	Support demand-controlled ventilation, zonal control and occupant-centered regulation

For control-oriented modeling, these data should be organized according to operational meaning. Weather and occupancy form disturbance inputs. Equipment operation forms control inputs. The indoor environment represents the system state. Energy use and comfort define objectives and constraints. Short-term energy prediction generally depends on historical load, time features and weather forecasts. Indoor temperature prediction additionally requires information on thermal inertia and heating or cooling supply. Demand-controlled ventilation depends on occupancy intensity and CO_2 accumulation. Personalized comfort control further requires the linkage between local environmental conditions and user preferences. The key issue is therefore not the number of input variables, but whether the model can represent the dynamic mapping among disturbances, control actions and environmental responses.

2.2. Prediction of energy use and HVAC loads

HVAC load prediction connects operational sensing with active control. Prediction outputs can support pre-cooling and pre-heating, equipment scheduling, demand response and MPC rolling optimization. Their usefulness depends strongly on the consistency between prediction horizon and application scenario. Minute-level and hourly prediction are closer to real-time control requirements, whereas daily and monthly prediction mainly support operation management, energy auditing and retrofit evaluation. Recent reviews show that prediction targets have expanded from whole-building energy use to cooling and heating loads, equipment-level energy consumption and multi-zone load

DOI: 10.54254/2755-2721/2026.34152

distribution [4]. Research attention has also shifted from model comparison alone to data quality, feature construction, prediction horizon and generalization capability.

Conventional machine learning models remain competitive in practical buildings. Support vector regression is suitable for nonlinear prediction tasks with limited samples. Random forests and gradient boosting can process multi-source features and provide partial interpretation of variable contributions. Shallow neural networks are commonly used to approximate nonlinear relationships among weather, time and equipment variables. With the accumulation of high-frequency sensor data, LSTM, CNN-LSTM and Transformer-based models have been introduced for multivariate time-series prediction. They can capture periodicity, lag effects and long-term dependencies. Existing studies suggest that deep models perform well when data volume and operating-condition coverage are sufficient, while simpler models can be more stable under missing data, limited samples or seasonal transfer [3, 4].

For control-oriented applications, the evaluation of load prediction models must extend beyond offline MAE or RMSE. Multi-step error accumulation, robustness under extreme weather, prediction uncertainty and online updating all affect the controller's assessment of future states. In MPC, prediction errors may be amplified by the optimizer into inappropriate control actions and may further cause comfort or equipment constraints to be violated. As a result, load prediction research is moving from static error comparison toward operational reliability assessment. Online learning-enhanced control feeds recent operational data back to the prediction model and continuously corrects model parameters to mitigate operating-condition drift [7]. This indicates that load prediction is evolving from an independent offline regression task into a dynamic state-estimation component within building control systems.

2.3. Thermal environment, air quality and comfort prediction

HVAC control is constrained by both energy objectives and indoor environmental requirements. Indoor temperature and humidity determine thermal deviation. CO_2 concentration reflects ventilation demand and air-quality risk. Thermal comfort indices connect physical environmental conditions with human perception. Compared with load prediction, thermal-environment and comfort prediction more directly influence the acceptability of control strategies. Energy-saving measures are meaningful only when basic comfort and health requirements are maintained.

Traditional thermal comfort assessment is represented by the PMV/PPD framework. It incorporates air temperature, relative humidity, mean radiant temperature, air velocity, clothing insulation and metabolic rate to evaluate average thermal sensation at the group level. This framework has a fundamental role in environmental design and standards, but its group-average nature limits its applicability in personalized control. In real buildings, thermal sensation is affected by age, gender, activity level, clothing habit, thermal history and psychological expectation. Different occupants may report different thermal preferences under the same environmental conditions. Fixed temperature setpoints are convenient in engineering practice, but they may not fully adapt to diverse users and dynamic space-use patterns.

Data-driven thermal comfort models incorporate environmental measurements, human states and subjective feedback to predict thermal sensation votes, thermal preferences or comfort probabilities. Environmental measurements provide the physical basis. Human-state information enhances the representation of individual differences. Subjective feedback supplies direct supervision signals. Feng et al. [8] systematically reviewed data-driven personal thermal comfort prediction and identified data collection, model construction and application scenarios as key issues. Compared with conventional PMV-based assessment, data-driven methods can adapt more flexibly to different

DOI: 10.54254/2755-2721/2026.34152

spaces, populations and operating modes, and can support zonal control, local air distribution and personalized setpoint adjustment.

The application of personal comfort models is still limited by data acquisition and cross-context generalization. Subjective feedback requires long-term occupant participation, and frequent requests may reduce user acceptance. Image recognition, wireless positioning and wearable devices can provide richer human-state information, but they also raise privacy, cost and data-governance concerns. Models trained for a specific building or population may suffer performance degradation across seasons, zones or user groups. Avci [9] introduced federated learning for real-time zone-based personalization of indoor thermal comfort, which provides a path for privacy-preserving and cross-zone modeling. Thermal comfort prediction should therefore move from classification accuracy toward low-intrusion sensing, privacy protection, model transfer and control-interface design.

2.4. Occupant behavior modeling and its influence on predictive control

Occupant behavior affects load formation, environmental demand and control acceptance. Presence status, staying time and activity level change internal sensible and latent loads as well as ventilation demand. Window operation, shading adjustment, lighting use and thermostat changes alter envelope heat exchange and equipment load. Comfort feedback reveals the deviation between control strategies and user expectations. Xu et al. [5] showed that occupant behavior research has moved from static schedule assumptions toward dynamic behavior modeling based on multi-source sensing, probabilistic methods and control integration.

In HVAC control, occupant information operates through several mechanisms. It changes the spatial and temporal allocation of comfort and ventilation constraints. Low-occupancy zones may allow wider temperature or ventilation bands, while high-occupancy zones require stricter thermal and air-quality protection. It also changes short-term load prediction because occupant density, activity intensity and behavioral interventions affect internal loads and equipment responses. In addition, occupant preference affects the acceptability of control strategies. A numerical energy-saving strategy may be rejected in practice if it causes frequent comfort complaints or reduces users' sense of control.

Occupancy detection, occupant-count estimation, activity recognition, interaction-behavior prediction and comfort-feedback modeling form a multi-layered research system. Infrared sensing, access-control records, CO_2 concentration, Wi-Fi, cameras and plug-load patterns have all been used for occupancy inference. These technologies differ in accuracy, cost, spatial resolution and privacy sensitivity. Coarse occupancy information can improve whole-building load prediction, while zonal air distribution and local comfort control require finer information on regional population, activity and interaction behaviors. Esrafilian-Najafabadi et al. [10] showed that occupancy prediction models significantly influence both energy savings and thermal comfort when incorporated into HVAC control. This finding supports the evaluation of occupancy modeling within a control-performance framework rather than as an isolated classification task.

3. Data-driven HVAC control for coordinated comfort and energy efficiency

Building operation optimization requires predictive information to be converted into executable HVAC control actions. Data-driven control uses weather forecasts, occupancy status, indoor environmental states and equipment operation data to evaluate how alternative control strategies affect energy use, comfort and indoor air quality. The controller can then adjust setpoint temperature, air flow rate, supply-water temperature, fan frequency or damper opening. Compared

DOI: 10.54254/2755-2721/2026.34152

with offline prediction, control-oriented research emphasizes closed-loop stability, constraint satisfaction, real-time computation and deployability.

3.1. From reactive control to predictive control

Traditional HVAC control is mainly based on scheduled operation, fixed setpoints, PID control and empirical rules. These methods have clear structures and are convenient for operation and maintenance. Their limitation is associated with the reactive adjustment mechanism. Control actions are usually generated from the deviation between current measurements and setpoints, with limited use of future weather, occupancy changes, load fluctuations and electricity-price information. Buildings with strong thermal inertia may suffer from delayed response. Spaces with fluctuating occupancy may experience unnecessary conditioning during vacant periods. Demand-response scenarios also require stronger prediction capability to shift loads proactively.

MPC enhances the use of future disturbances through receding-horizon optimization. At each control step, the prediction model estimates indoor temperature, humidity, CO_2 concentration and equipment energy use over a future horizon. The optimizer then solves the control sequence under comfort, air-quality and equipment constraints. Data-driven models reduce the detailed modeling and parameter-calibration burden of white-box MPC. They also allow the controller to learn response characteristics directly from operational data. The value of data-driven MPC arises from the coupling among prediction, constraints and rolling feedback. The prediction model estimates future states, the constraints define comfort and safety boundaries, and feedback updates subsequent control actions using recent measurements.

3.2. Coupling between predictive models and optimizers in data-driven MPC

Data-driven MPC requires predictive models that can be stably called by the optimizer and can respond reasonably to changes in control variables. A model that only forecasts natural system evolution is insufficient for control optimization, because the optimizer must compare the effects of different setpoints, air-flow rates or damper positions on future indoor states and energy use. A control-oriented predictive model should therefore represent the relationship among external disturbances, control actions and system responses.

Stoffel et al. [11] compared artificial neural networks, Gaussian process regression and linear regression in real-life MPC applications in an office and a test hall. The neural network showed better offline prediction accuracy, but the differences in actual control performance among the models were not significant. The linear model was more practical in deployment because of its simple structure, low tuning effort and stronger interpretability. This result indicates that model evaluation for control should consider prediction accuracy together with computational efficiency, optimization stability, parameter calibration and operational interpretability.

Sha et al. [7] used an encoder-decoder LSTM to predict future HVAC energy use, CO_2 concentration, air temperature and relative humidity. Model parameters were continuously updated according to operational errors. At the control level, multi-objective particle swarm optimization searched for feasible solutions, while adaptive weighting coordinated energy use, indoor air quality and thermal comfort. This framework extends single-output energy prediction to multi-state prediction and embeds model correction into rolling control. It illustrates the transition from offline modeling to online optimization.

Equipment-level studies further show how data-driven control can be connected to physical actuators. Jing et al. [12] established a multilayer perceptron model for a multi-zone demand-

DOI: 10.54254/2755-2721/2026.34152

controlled ventilation system. The model predicted damper pressure differences from main-duct static pressure and target air flow, and damper openings were then determined using empirical damper equations. The method mitigated over-ventilation and under-ventilation and showed fan energy-saving potential. Such studies extend data-driven control from whole-building optimization to detailed regulation of equipment actions and air-flow distribution.

3.3. Multi-objective optimization, demand-controlled ventilation and reinforcement learning

HVAC optimal control is a dynamic decision problem with multiple objectives and constraints. Reducing cooling, heating or ventilation rates can lower energy use but may increase thermal discomfort or air-quality risks. Maintaining strict temperature, humidity and ventilation requirements improves environmental quality but raises equipment load and operating cost. Thermal comfort can be represented by temperature range, PMV constraint, comfort deviation or user satisfaction. Air quality is often expressed by CO_2 or pollutant concentration. Energy objectives can be described by HVAC power, cumulative energy use, operating cost or carbon emission factor. Equipment constraints include fan-frequency limits, damper-opening limits, chiller start-stop frequency, water-temperature range and control-action variation.

Occupant information changes the spatial and temporal weighting of the optimization problem. Low-occupancy zones can operate with wider temperature or ventilation bands, while high-occupancy zones require priority protection of comfort and air quality. For short-use spaces, pre-conditioning and thermal-inertia utilization can reduce the energy penalty associated with continuously maintaining fixed setpoints. Occupant acceptance also matters.

Multi-zone demand-controlled ventilation and variable-air-volume systems are typical scenarios for coordinated optimization. Different zones may have different occupant densities, CO_2 concentrations and thermal loads. Uniform ventilation or temperature settings can therefore cause local over-ventilation or under-ventilation. Shi et al. [13] proposed an event-driven optimal demand-controlled ventilation method that combined an artificial neural network model of the ventilation network with a multi-zone thermal network model. The method directly optimized fan frequency and damper opening and used event triggering to reduce unnecessary optimization. This type of research demonstrates that HVAC optimization must consider environmental quality, energy efficiency, computation load and deployment feasibility at the same time.

Reinforcement learning provides an adaptive optimization route based on policy learning. It can represent indoor environment, equipment state, weather and occupancy as states, use setpoint temperature, air flow rate, damper opening or equipment on/off status as actions, and include energy use, comfort, air quality and cost in the reward function. Simulation studies have demonstrated its energy-saving potential, but deployment in real buildings is still constrained by safe exploration, reward-function design, simulation-to-reality mismatch and limited interpretability [14, 15]. A feasible near-term pathway is to combine reinforcement learning with rule-based control, MPC, safety constraints and manual override mechanisms. Its long-term value depends on whether policy learning can be made safe, transferable and interpretable in operational buildings.

4. Conclusions and outlook

This paper reviewed data-driven methods for thermal-environment regulation and HVAC energy optimization in smart buildings. The discussion covered multi-source operational data, energy and load prediction, indoor environmental and thermal comfort prediction, occupant behavior modeling and HVAC optimal control. Current research is moving from single energy-series prediction toward

DOI: 10.54254/2755-2721/2026.34152

multi-state joint modeling. Load prediction provides feedforward information for equipment scheduling and MPC. Thermal-environment and air-quality prediction support control constraints. Occupant behavior modeling links energy supply with actual space-use demand.

From the control perspective, data-driven methods are promoting a transition from reactive HVAC operation toward predictive and adaptive regulation. Data-driven MPC uses operational data to construct system-response models and coordinates energy use, thermal comfort, air quality and equipment constraints over a rolling horizon. Real-building studies indicate that offline prediction accuracy alone cannot represent closed-loop control performance. Model structure, optimization stability, online updating and operational interpretability also determine deployment quality [7, 11]. Demand-controlled ventilation and multi-zone VAV studies further show that data-driven methods can support detailed regulation of fan, damper and fresh-air operations [12, 13]. Reinforcement learning offers a strategy-learning route for adaptive control, while its real-building use remains limited by safety, transferability and explainability [14, 15].

Future research can be advanced in four directions. First, physical constraints should be integrated more deeply with data-driven models. Embedding thermal balance, equipment performance curves, comfort constraints and control boundaries can improve reliability under small samples, extreme weather and extrapolated conditions. Second, thermal comfort and occupant modeling should move toward personalization and privacy protection. Federated learning and other privacy-preserving methods can help address data-sharing restrictions while retaining the advantages of occupant-centered modeling [9]. Third, model evaluation should extend from offline error metrics to closed-loop performance, including energy savings, comfort compliance, IAQ, equipment-action smoothness, constraint violations and user acceptance. Fourth, transferability and system integration should be strengthened. Differences in climate, building function, equipment type and occupant behavior make direct model transfer difficult. Transfer learning, domain adaptation, online learning and standardized data interfaces can reduce the cost of cross-building deployment.

Overall, data-driven methods provide important support for low-carbon HVAC operation and smart building environmental regulation. Their future value will depend on reliability, interpretability and deployability in real operating conditions. A closer integration of building physics, multi-source operational data, occupant demand and control-system constraints is required to reduce energy use while maintaining indoor environmental quality.

References

[1] International Energy Agency. Buildings. IEA, 2023.
[2] Global Alliance for Buildings and Construction, United Nations Environment Programme. Global Status Report for Buildings and Construction 2024/25. UNEP, 2025.
[3] Balali Y, Chong A, Busch A, O'Keefe S. Energy modelling and control of building heating and cooling systems with data-driven and hybrid models: A review. Renewable and Sustainable Energy Reviews, 2023, 183: 113496.
[4] Li Y, Arellano-Espitia F, Aler R, Igualada L, Corchero C. Data-driven methods and their applications to building HVAC energy consumption prediction: A review. Journal of Building Engineering, 2025, 116: 114612.
[5] Xu X, Yu H, Sun Q, Tam V W Y. A critical review of occupant energy consumption behavior in buildings. Renewable and Sustainable Energy Reviews, 2023, 182: 113396.
[6] Huang G, Ng S T, Li D, Zhang Y. State of the art review on the HVAC occupant-centric control in different commercial buildings. Journal of Building Engineering, 2024, 96: 110445.
[7] Sha X, Ma Z, Sethuvenkatraman S, Li W. Online learning-enhanced data-driven model predictive control for optimizing HVAC energy consumption, indoor air quality and thermal comfort. Applied Energy, 2025, 383: 125341.
[8] Feng Y, Ouyang Q, Zhu Y, et al. Data-driven personal thermal comfort prediction: A literature review. Renewable and Sustainable Energy Reviews, 2022, 161: 112357.

DOI: 10.54254/2755-2721/2026.34152

[9] Avci A B. Federated machine learning for indoor thermal comfort: Real-time zone-based personalization and HVAC control. Journal of Building Engineering, 2025, 114: 114476.
[10] Esrafilian-Najafabadi M, Haghighat F, et al. Impact of occupancy prediction models on building HVAC control system performance. Energy and Buildings, 2022.
[11] Stoffel P, Berktold M, Müller D. Real-life data-driven model predictive control for building energy systems comparing different machine learning models. Energy and Buildings, 2024, 305: 113895.
[12] Jing G, Cai W, Zhang X, Cui C, Liu H, Wang C. An energy-saving control strategy for multi-zone demand controlled ventilation system with data-driven model and air balancing control. Energy, 2020, 192: 116714.
[13] Shi S, Miyata S, Akashi Y. Event-driven model-based optimal demand-controlled ventilation for multizone VAV systems. Applied Energy, 2025, 377: 124683.
[14] Al Sayed K, Boodi A, Sadeghian Broujeny R, Beddiar K. Reinforcement learning for HVAC control in intelligent buildings. Journal of Building Engineering, 2024, 95: 110085.
[15] Kannari L, Wessberg N, Hirvonen S, Kantorovitch J, Paiho S. Reinforcement learning for control and optimization of real buildings. Journal of Building Engineering, 2025, 104: 112283.