



Research paper

Prediction-guided deep reinforcement learning for energy-efficient HVAC control with PICVs in smart buildings

Younghyun Baek^a, Guwon Yoon^b, Seunghwan Kim^{a,b}, Byeongkwan Kang^a,
Sanghoon Lee^a, Myeong-in Choi^a, Sehyun Park^{a,b,*}

^a Department of Intelligent Energy and Industry, Chung-Ang University, Seoul 06974, Republic of Korea

^b School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Republic of Korea

ARTICLE INFO

Keywords:

Deep reinforcement learning
Smart HVAC control
Thermal comfort
Time series forecasting
Data-driven control strategies

ABSTRACT

This study addresses a structural limitation of existing HVAC control frameworks by introducing a prediction-grounded, physics-engine-free, data-driven formulation in which system dynamics are approximated using data-driven forecasts and pressure-independent control valves (PICVs) are treated as controllable optimization variables. The framework integrates a Gated Recurrent Unit (GRU) time-series forecaster with a Proximal Policy Optimization (PPO) agent, injecting look-ahead forecasts into the agent's state. The GRU model predicts short-horizon HVAC energy and indoor conditions, and these look-ahead signals condition the PPO state to enable anticipatory control. Using these forecasts, the agent learns continuous policies for Air Handling Unit (AHU) fan speeds and PICV openings to minimize energy use under an explicit Predicted Mean Vote (PMV)-aware reward that constrains comfort within the target range. Experiments with real-world data from July to September demonstrate up to 11.85% energy savings relative to conventional control. PMV remained within the ANSI/ASHRAE 55 comfort band ($PMV \in [-0.5, 0.5]$) during operating periods, evidencing comfort preservation alongside energy savings. These findings highlight fine-grained, anticipatory control and consistent energy performance gains. The proposed framework is scalable to multi-zone HVAC control and provides a foundation for potential extension toward district- and city-scale energy management. These results demonstrate that prediction-grounded, physics-engine-free control provides a robust data-driven foundation for anticipatory HVAC control in real operational buildings.

1. Introduction

Climate change and global warming are among the most critical challenges facing humanity. In response, the global community has emphasized carbon-emission reductions through policies and technologies aimed at carbon neutrality. The building sector alone accounts for roughly 40% of global energy use and a comparable share of carbon emissions, making its optimization a priority for sustainable development (IEA, 2024; Ocampo Batlle et al., 2020). Broadly, two strategies are pursued: (i) optimizing energy production and integrating renewables, and (ii) reducing consumption in existing buildings through operational optimization (Choi et al., 2024). This study adopts the latter, focusing on HVAC systems — the dominant end-use in buildings, often exceeding 40% of building energy consumption — whose optimization is central to emission reduction and is a focal point of current policy and technical roadmaps (Dal Cin et al., 2022). Advances in information and communication technology (ICT) and Internet of Things

(IoT) instrumentation enable the *smart building* paradigm, in which high-resolution operational data are collected for real-time analytics and optimization. Compared with purely physics-based simulations, data-driven approaches can capture operational variability and unmodeled dynamics in real buildings, improving forecasting fidelity and control robustness (Amber et al., 2018). Prior work spans classical time-series models (e.g., autoregressive (AR) and autoregressive integrated moving average (ARIMA)) and deep models such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and temporal convolutional networks (TCNs) (Chou and Tran, 2018; Chaturvedi et al., 2022; Li et al., 2023; Somu et al., 2021; Akter et al., 2025). While these studies have improved prediction accuracy and informed control design, much of the control-side literature still relies on physics engines (e.g., EnergyPlus, TRNSYS) for evaluation (Nguyen et al., 2024; Fu et al., 2023). In contrast, we investigate a prediction-guided data-driven control framework in a multi-zone building within

* Corresponding author.

E-mail addresses: yhbaek@cau.ac.kr (Y. Baek), gw1206@cau.ac.kr (G. Yoon), tkftn456@cau.ac.kr (S. Kim), byeongkwan@cau.ac.kr (B. Kang), leessan0@cau.ac.kr (S. Lee), auddiscjswo@cau.ac.kr (M.-i. Choi), shpark@cau.ac.kr (S. Park).

<https://doi.org/10.1016/j.egyr.2026.109040>

Received 14 November 2025; Received in revised form 16 January 2026; Accepted 24 January 2026

Available online 29 January 2026

2352-4847/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

Nomenclature

HVAC	Heating, Ventilation, and Air Conditioning	FCU	Fan Coil Unit
PICVs	Pressure-Independent Control Valves	DRL	Deep Reinforcement Learning
GRU	Gated Recurrent Unit	DNN	Deep Neural Network
PPO	Proximal Policy Optimization	MDP	Markov Decision Process
AHU	Air Handling Unit	MPC	Model Predictive Control
PMV	Predicted Mean Vote	SAC	Soft Actor-Critic
ICT	Information and Communication Technology	FA-MADRL	Federated Accelerated Multi-Agent Deep Reinforcement Learning
IoT	Internet of Things	Seq2Seq	Sequence-to-Sequence
AR	Autoregressive	CNN	Convolutional Neural Network
ARIMA	Autoregressive Integrated Moving Average	SVR	Support Vector Regression
RNN	Recurrent Neural Network	GRNN	General Regression Neural Network
LSTM	Long Short-Term Memory	A2C	Advantage Actor-Critic
TCN	Temporal Convolutional Network	GAE	Generalized Advantage Estimation
TSF	Time Series Forecasting	RMSE	Root Mean Square Error
ANSI	American National Standards Institute	SMAPE	Symmetric Mean Absolute Percentage Error
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers	PIR	Passive Infrared
VFD	Variable Frequency Drive	ESS	Energy Storage System
BEMS	Building Energy Management System	PV	Photovoltaics

a multi-building complex. Specifically, we integrate a GRU-based time-series forecasting (TSF) model with a Proximal Policy Optimization (PPO) agent. The GRU forecaster provides short-horizon look-ahead signals of HVAC energy and indoor conditions that are injected into the agent's state, enabling anticipatory (rather than purely reactive) control.

The agent then learns continuous control policies for AHU fan speeds and PICVs opening rates to minimize energy consumption while preserving thermal comfort via a PMV-aware reward that enforces the ANSI/ASHRAE 55 comfort band ($PMV \in [-0.5, 0.5]$) during operating periods. Experiments using real building data from July to September show up to 11.85% energy savings relative to baseline (RBC), while maintaining PMV within the comfort range across operating periods. The proposed framework is readily scalable to multi-zone HVAC control and provides a practical foundation for potential extension toward district- and city-scale coordination.

Contributions. This study proposes a prediction-grounded, physics-engine-free formulation of HVAC control in which data-driven forecasts explicitly represent system dynamics and PICVs are treated as controllable optimization layers. The main contributions of this study are summarized as follows:

- **Prediction-grounded, physics-engine-free anticipatory HVAC control:** We propose a physics-engine-free, prediction-grounded deep reinforcement learning framework in which a GRU-based forecasting model explicitly serves as a learned transition surrogate of short-horizon system dynamics. The predicted indoor environmental states and energy-related variables are directly injected into the PPO agent's state, enabling anticipatory control that compensates for unmodeled disturbances and short-term uncertainties (e.g., occupant behavior, intermittent door openings, and rapid indoor-outdoor thermal interactions) that cannot be captured by rule-based or purely reactive control schemes.
- **Control paradigm shift beyond algorithm substitution:** Rather than replacing one learning algorithm with another, this study demonstrates a shift from reactive HVAC control toward prediction-informed decision making. The GRU forecaster is not used solely to improve prediction accuracy, but to provide look-ahead state information that fundamentally alters the control logic of the DRL agent. This distinguishes the proposed GRU-PPO framework from existing prediction-assisted control approaches that rely on simplified dynamics or indirect forecast utilization.

• PMV-constrained continuous control using real building data:

We design a PMV-aware reward function that explicitly enforces thermal comfort constraints according to ANSI/ASHRAE 55 ($PMV \in [-0.5, 0.5]$) while optimizing energy consumption. The proposed framework jointly controls continuous AHU fan speeds (via VFDs) and floor-level PICV opening rates using high-resolution IoT/BEMS data collected from a multi-building, multi-zone environment. Experimental results demonstrate up to 11.85% energy savings compared to a rule-based baseline, while maintaining indoor comfort within the prescribed bounds.

- **Practical scalability and real-world relevance:** By relying on measured operational data rather than physics-based simulators, the proposed framework reflects real building behavior and operational variability. The control structure is readily extensible to larger multi-zone and multi-building systems, providing a practical pathway toward district- and city-scale HVAC energy management.

2. Related work

Predictive deep reinforcement learning (DRL) has been widely investigated as a promising approach to simultaneously improve HVAC energy efficiency and indoor thermal comfort. A substantial body of work has focused on forecasting indoor environmental and energy variables, because accurate short-horizon predictions can serve as informative signals for control. Representative studies include multi-zone indoor temperature prediction using LSTM-based sequence-to-sequence (Seq2Seq) models (Fang et al., 2021; Minassian et al., 2025) and hybrid CNN-LSTM architectures that capture both spatial and temporal dependencies for predicting temperature, humidity, and comfort-related indices (Elmaz et al., 2021). Other studies employed LSTM to predict energy consumption and indoor temperature dynamics across different temporal resolutions (Shakeel et al., 2023; Sendra-Arranz and Gutiérrez, 2020; Xiao et al., 2022), and comparative evaluations across RNN-, LSTM-, and GRU-based forecasters have reported that GRU can provide competitive accuracy with reduced complexity for building time series (Taylor and Letham, 2018). Beyond the core deep models, recent studies explored forecasting performance improvements via preprocessing and modeling choices or by benchmarking alternative predictors, including thermal/energy forecasting (Peng et al., 2023), heating operation pattern prediction (Jang et al., 2022),

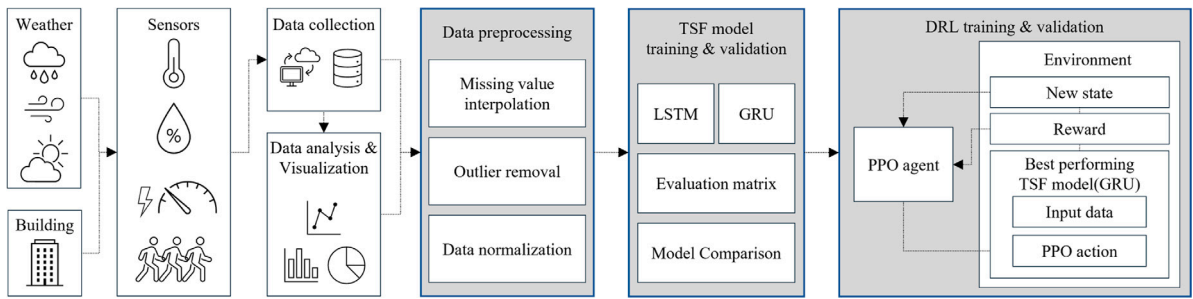


Fig. 1. Overall structure of the proposed TSF-PPO-based HVAC optimization framework.

hyperparameter-focused deep learning studies (Morteza et al., 2023), evolutionary/ensemble-style learning for energy prediction (Almalaq and Zhang, 2019), household-level hybrid forecasting (Syed et al., 2021), and long-term techno-economic analyses leveraging LSTM/GRU prediction results (Qays et al., 2025). Collectively, these studies confirm that (i) indoor-state forecasting is feasible and beneficial, but (ii) real buildings still exhibit unmodeled disturbances and short-term uncertainties that are difficult to capture using only measured exogenous variables.

In parallel, reinforcement learning-based HVAC control has evolved from early value-based methods, such as Q-learning for energy/comfort optimization (Chen et al., 2018), toward actor-critic and policy-gradient methods that better support continuous control (Biemann et al., 2021). Further extensions include hierarchical control frameworks (Wang and Wang, 2021), cost-aware MPC formulations (Bird et al., 2022), and distributed/federated reinforcement learning for large-scale systems (Xia et al., 2024). Several studies additionally considered human/occupant interactions and adaptive operation policies (Baldi et al., 2018). Importantly, prediction-guided RL has been explored by combining an LSTM predictor with a continuous-control RL agent (e.g., SAC) (Zhuang et al., 2023), showing that injecting look-ahead information can improve performance over purely reactive control. However, many RL studies are still evaluated in simplified or idealized settings, and safety/reliability concerns remain a key barrier to deployment; for example, batch reinforcement learning with safety constraints has been proposed for building HVAC to mitigate unsafe exploration risks (Zhang et al., 2022). Meanwhile, conventional rule-based control (RBC) continues to be widely used in practice due to transparency and operational robustness. Recent work has revisited rule-based approaches by enhancing them with lightweight intelligence, including fuzzy logic-based optimization for air-conditioning energy reduction (Pagare et al., 2025) and machine learning-enhanced lightweight RBC strategies that aim for practical demand-response or energy savings with limited complexity (Zhu et al., 2025); additionally, rule extraction has been studied to improve the interpretability of deep RL controllers (Razzano et al., 2025). Although these RBC or lightweight schemes are attractive for buildings with limited sensing or compute, their performance is inherently bounded by predefined rules and often under-utilizes rich indoor sensing data when available. Moreover, most RL-based HVAC controllers remain fundamentally dependent on physics-based simulators and predominantly adopt reactive formulations, which limits robustness under real operational disturbances.

Moreover, existing HVAC control studies predominantly focus on AHU-, FCU-, or zone-temperature-level actuation, while pressure-independent control valves (PICVs) are generally treated as passive hydraulic components. Their potential as controllable optimization layers in data-driven reinforcement learning frameworks remains largely unexplored.

Overall, prior studies remain limited by (i) simulator-bound and reactive HVAC control formulations, (ii) vulnerability to short-term real-building disturbances, and (iii) the absence of PICV-layer actuation as a controllable optimization dimension. Motivated by these

gaps, we propose a prediction-grounded, physics-engine-free GRU-PPO framework that embeds short-horizon forecasts as representations of system dynamics and jointly optimizes AHU and PICV control under PMV-constrained comfort requirements.

3. Methodology

In this study, a DRL-based approach was employed to optimize the HVAC system. The target of this research is a specific building in a complex featuring a multi-building HVAC system equipped with an IoT sensor network that collects various environmental data in real-time, providing a suitable environment for implementing a data-driven HVAC control strategy. Leveraging this infrastructure, the study proposes an optimal HVAC operation plan that considers both energy efficiency and occupant comfort. Occupant comfort was assessed using the PMV index, widely adopted in existing standards such as ANSI/ASHRAE 55-2020 (Yang et al., 2014), which specifies an acceptable PMV range of -0.5 to $+0.5$. This study quantified comfort by accounting for factors such as air temperature, mean radiant temperature, relative humidity, airflow velocity, clothing insulation, and metabolic rate. A DRL-based control model was trained to minimize energy consumption while maintaining comfort within the specified PMV range (ASHRAE, 0000). The target building collects indoor and outdoor environmental data such as temperature, humidity, wind speed, and CO_2 concentration via IoT sensors, while HVAC-related information is obtained through the Building Energy Management System (BEMS). In addition, the AHUs in the building are configured to operate in a frequency control mode rather than a simple on/off method and to independently regulate the flow rate of chilled water supplied to the fan coil units (FCUs) through PICVs installed on each floor. This system configuration provides a suitable structure for reinforcement learning-based control, with PICVs recognized as key components contributing to improved energy efficiency (Naldi et al., 2023). In this study, the HVAC control policy was trained and evaluated in a physics-engine-free, data-driven simulation environment driven by measured IoT/BEMS operational data, rather than being deployed in real-time closed-loop control.

This approach enabled an evaluation of the effectiveness of DRL-based policies under conditions close to real-world operating environments and indicated the potential for future extension to real-time control (see Figs. 1 and 3).

After collecting the environmental and equipment data, missing values and outliers were addressed and the data were normalized. A time series forecasting (TSF) model was subsequently trained, and the performance of LSTM- and GRU-based models was compared. The best-performing model was then selected and used as the state input for the DRL environment. DRL defines HVAC control variables, specifically the AHU fan speed and the PICV opening rate, as actions and learns optimal control policies based on the environmental state predicted by the TSF model. An equation was formulated for the multi-building HVAC system to divide the total energy consumption by the building-specific energy consumption. In this study, the optimization specifically targets the energy consumption of the HVAC system within a single building,

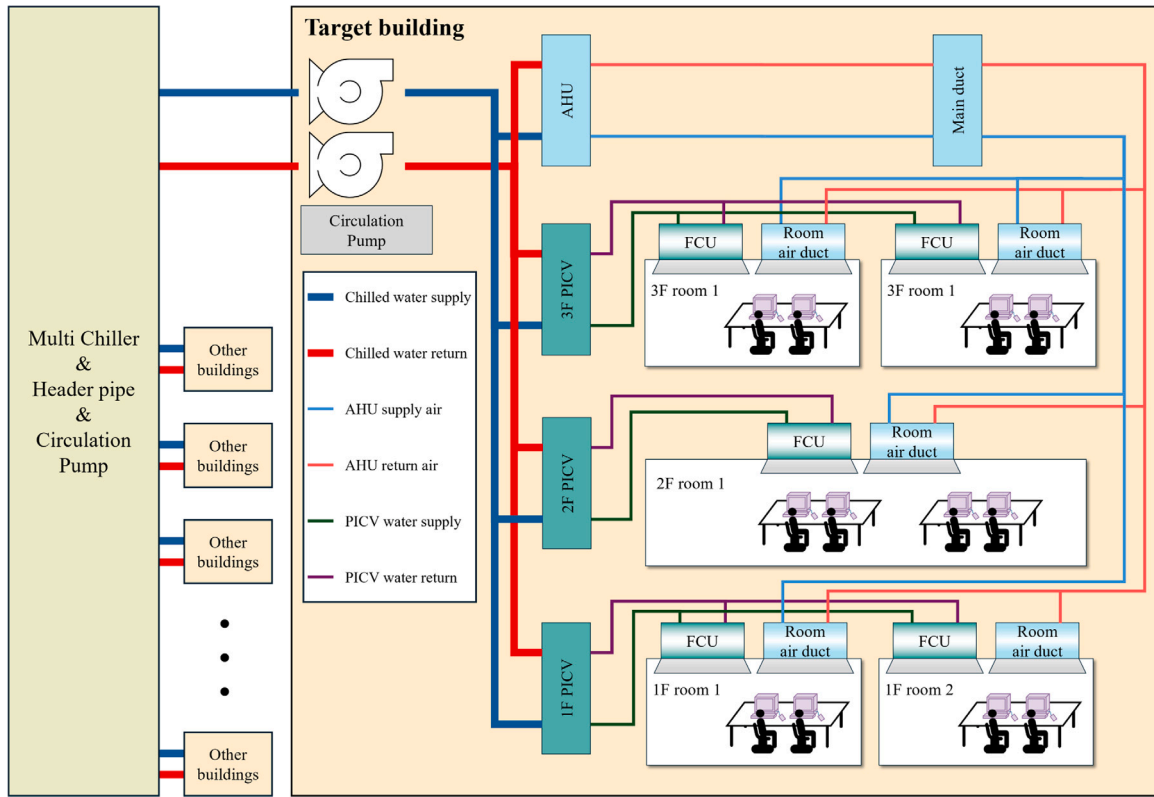


Fig. 2. Chiller and chilled water circulation system for the multi-building HVAC operation.

excluding the energy usage of personal computers and other electric appliances operated by occupants. The total energy consumption of the HVAC system in the building is calculated as the sum of the following three components:

$$E_{AHU} = E_{supply\ fan} + E_{return\ fan} \quad (1)$$

$$E_{Heat} = Chilled\ water\ flow\ rate \times \Delta Chilled\ water\ temperature \quad (2)$$

$$E_{HVAC} = E_{AHU} + E_{Heat} + E_{Circulation\ pump} \quad (3)$$

The formulas above represent the energy consumption of the HVAC system optimized in this study, and thermal energy calculations were employed to separately measure the HVAC energy consumption of each building within the multi-building environment (Yu et al., 2024).

3.1. Prediction targets and input features

The time-series forecasting (TSF) task was formulated as a multi-output prediction problem. The target variables include HVAC system energy consumption and indoor thermal conditions (air temperature and relative humidity) across multiple floors and rooms. Specifically, one energy-related variable and multiple zone-level environmental variables were jointly predicted. All remaining measured variables, including HVAC operational data (e.g., AHU fan speed and PICV opening rate), indoor environmental measurements, occupancy-related signals, and outdoor weather conditions, were used as exogenous input features. All variables were normalized using Min–Max scaling to ensure numerical stability during model training. The dataset was divided into training and testing sets with an 80:20 ratio while preserving the temporal order of the time series.

3.2. Physics-engine-free, data-driven simulation environment

The multi-building complex in which the target building is located operates eight chillers: four turbo chillers and four absorption chillers. Each chiller is equipped with a chilled water pump and a cooling tower. The system also includes chilled water pumps and chilled water headers, which supply chilled water from the chillers and collect it after use. Chilled water is essential for the operation of the chillers; the supplied chilled water absorbs heat during usage and is then returned. After absorbing heat, the used cooling water is pumped to the cooling towers via cooling water pumps, where it is cooled and recirculated as supply water. In the system, the chilled water produced by the chillers is collected in the chilled water header and distributed through the chilled water pipes by four primary circulation pumps. The water circulates throughout the entire building, absorbing heat, and is then returned to the chillers via the return water pipes. As the chilled water circulates throughout the building, the secondary circulation pumps draw the required amount of water for each building, and the used water is routed back through the return pipes. The water, having increased in temperature, reenters the chillers, where it is recooled and recirculated.

Fig. 2 illustrates the multi-chiller system, including the header pipes and primary circulation pump, represented as a box. This system delivers chilled water to the target building, which is depicted in the rightmost box. The chilled water received from the primary circulation pump is supplied to the areas that require cooling via the secondary circulation pump. This chilled water is distributed to the AHUs and FCUs. Before being supplied to the FCUs, the flow is controlled by PICVs on each floor, and the chilled water is collected at a single point via return pipes. The collected water is then directed to the multi-chiller unit through the primary circulation pump. In this study, data were collected over a six-month period, specifically during times of peak energy consumption and power demand, when the multi-chiller system was in operation. Using the flow rate and temperature difference of

the chilled water, the heating and cooling energy consumption for each building can be calculated.

3.2.1. Sequence construction

To capture temporal dependencies and system dynamics, the input sequence at time step t was constructed by concatenating both past exogenous inputs and past target variables over a fixed sequence length k . Formally, the input sequence is defined as

$$\mathbf{z}_{t-k:t-1} = [\mathbf{x}_{t-k:t-1}, \mathbf{y}_{t-k:t-1}],$$

where $\mathbf{x}$ denotes the exogenous input variables and $\mathbf{y}$ represents the target variables. The model then predicts the target vector $\mathbf{y}_t$ at the next time step. This formulation allows the forecasting model to learn the underlying environmental dynamics of the HVAC system by explicitly incorporating historical system responses, thereby providing a reliable approximation of system behavior for downstream control.

3.3. Long Short-Term Memory (LSTM)

This study employed LSTM as a time series-based prediction model to predict energy consumption and indoor environmental variables of HVAC systems and subsequently trained a reinforcement learning control model based on the predictions. LSTM, an extension of the RNN architecture, incorporates a gate-based structure to mitigate the issues of gradient vanishing and explosion, effectively handling long-term dependencies in data (Hochreiter and Schmidhuber, 1997; Hochreiter, 1998). It is particularly well-suited for time series prediction tasks because it overcomes the limitations of RNNs and effectively learns continuous time series data. LSTMs control the cell state through three gates: the forget gate, the input gate, and the output gate, which determine whether information is retained or updated. The computational process of an LSTM is given by:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \quad (6)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (7)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t * \tanh(C_t) \quad (9)$$

In addition, $\tilde{C}_t$, computed concurrently, represents the candidate cell state at the current time step, with values ranging between -1 and 1 after applying the $\tanh$ function. The previous cell state C_{t-1} combined with the output f_t of the forget gate, as well as the candidate cell state $\tilde{C}_t$ and the output i_t of the input gate, are used to compute the current cell state C_t . The output gate o_t determines which information from the current cell state C_t is passed to the covert state h_t . The final hidden state h_t is computed by applying the $\tanh$ function to C_t and then multiplying the result by o_t . This gate-based structure enables the LSTM to retain important information within the time series data over the long term while effectively filtering out irrelevant information.

In this study, a two-layer stacked LSTM architecture was implemented for time-series prediction. The first LSTM layer outputs the full hidden state sequence to preserve temporal information, while the second LSTM layer aggregates the temporal features into a compact latent representation. A fully connected layer with linear activation was used to generate multi-output predictions corresponding to the defined target variables. The LSTM model was trained using the Adam optimizer with mean squared error (MSE) as the loss function.

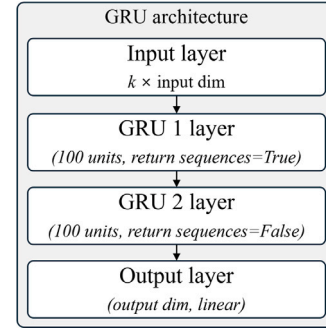


Fig. 3. Architecture of the stacked GRU model used for multi-output time-series forecasting.

3.4. Gated Recurrent Unit (GRU)

LSTMs mitigate the problems of vanishing and exploding gradients of RNNs by incorporating a gated architecture. However, this structural improvement increases computational complexity. To address this issue, the GRU, a simplified variant of the RNN, was introduced (Choi et al., 2014). The core idea of the GRU is to streamline the structure to enable more efficient time series prediction. While LSTMs manage the cell state and hidden state separately using three gates — input, forget, and output gates — the GRU consolidates the two states into a single unified state and relies on only two gates: the update gate and reset gate. The update gate regulates the extent to which new information is incorporated into the current state, and the reset gate controls whether information from the previous state is combined with the current input. Due to their simplified gate structure, GRUs require fewer parameters, leading to increased computational efficiency. In summary, LSTMs incorporate more parameters and sophisticated control structures to fine-tune the information flow, while GRUs provide fast and efficient learning with a simple structure. Thus, the choice of model largely depends on the nature of the data, its complexity, and computational resource constraints. In this study, LSTM and GRU models were trained as time series prediction models based on the same input structure, and their performance in predicting energy consumption and indoor environmental variables of HVAC systems was compared. As a result, the PPO algorithm was trained based on the model with better prediction accuracy to optimize the control policy of the HVAC system through reinforcement learning.

The GRU forecaster was implemented as a two-layer stacked GRU network followed by a linear output layer. Given an input sequence

$$\mathbf{z}_{t-k:t-1} \in \mathbb{R}^{k \times d_{\text{in}}},$$

where k denotes the sequence length and d_{in} represents the input feature dimension, the first GRU layer consists of 100 hidden units with `return_sequences=True`, enabling the output of the full hidden-state sequence to preserve temporal dependencies. The second GRU layer also contains 100 hidden units but is configured with `return_sequences=False` to generate a fixed-length latent representation summarizing the temporal information. This latent vector is then passed to a fully connected dense layer with linear activation, producing the multi-dimensional prediction

$$\hat{\mathbf{y}}_t \in \mathbb{R}^{d_{\text{out}}},$$

where d_{out} corresponds to the number of target variables. No dropout was applied in either GRU layer (dropout = 0). The model was trained using the Adam optimizer with mean squared error (MSE) as the loss function.

3.5. Deep reinforcement learning(DRL)

DRL integrates traditional reinforcement learning with Deep Neural Networks (DNNs) to enable effective policy learning in high-dimensional state spaces. This approach was first introduced by DeepMind Technologies, who demonstrated its scalability by using the Atari 2600 gaming environment to show that trained agents performed at the level of human experts. Since then, various researchers have refined DRL architecture and expanded its application to complex system control problems (Mnih et al., 2013). Reinforcement learning is typically defined as a Markov Decision Process (MDP), which consists of a state space S , an action space $\mathcal{A}$, a reward function $r : S \times \mathcal{A} \rightarrow \mathbb{R}$, a state transition probability $p(s'|s, a)$, and a discount factor $\gamma \in [0, 1]$. Within this framework, an agent selects an action a_t in a given state s_t , and optimizes a policy $\pi(a|s)$ while receiving a reward r_t and the next state s_{t+1} from the environment. The objective of reinforcement learning is to identify a policy π^* that maximizes the expected cumulative reward (Sutton and Barto, 1998).

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \quad (10)$$

In in-depth courses, neural networks are used to approximate state value functions $V^\pi(s)$, action value functions $Q^\pi(s, a)$, or policy functions $\pi(a|s)$. Notably, the actor-critic framework has been widely adopted in recent years to improve sample efficiency and stability. In this framework, the critic learns $Q^\pi(s, a)$ to evaluate values for state-action pairs, while the actor updates the policy $\pi(a|s)$ based on these values (Haarnoja et al., 2019).

$$Q^\pi(s, a) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right] \quad (11)$$

This approach is particularly effective in environments involving multidimensional inputs and nonlinear systems, such as HVAC control, and helps overcome the limitations of traditional model-based or rule-based control. On the other hand, model predictive control (MPC) enables precise control through optimization based on predictive models. However, due to computational complexity and model uncertainty in real-time operation, there has been growing interest in data-driven control based on DRL (Blum et al., 2022). In this study, we applied DRL to an HVAC system to achieve the dual objectives of reducing energy consumption while maintaining a stable comfort metric (PMV). To achieve this, a reinforcement learning agent learns real-time control strategies based on state data from the indoor environment and forms control policies that are adaptable to different operating conditions. DRL-based control can significantly improve control performance in smart building applications by achieving higher control precision and energy efficiency than conventional control.

3.6. Proximal Policy Optimization (PPO)

In this study, the PPO algorithm was adopted for reinforcement learning-based HVAC system control. PPO is a representative policy-based reinforcement learning algorithm designed to alleviate the instability and convergence problems of the existing Policy Gradient method (Wang et al., 2024). It builds on the Advantage Actor-Critic (A2C) framework and incorporates an objective function with a clipping mechanism to ensure learning stability and prevent performance degradation due to rapid changes in the policy update process (Mnih et al., 2016; Schulman et al., 2017). The surrogate loss function used by PPO is defined as follows:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (12)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ represents the probability ratio of the current policy to the previous policy, $\hat{A}_t$ denotes the advantage value calculated using Generalized Advantage Estimation (GAE), and ϵ is a hyperparameter that limits the amount of allowable policy change. This

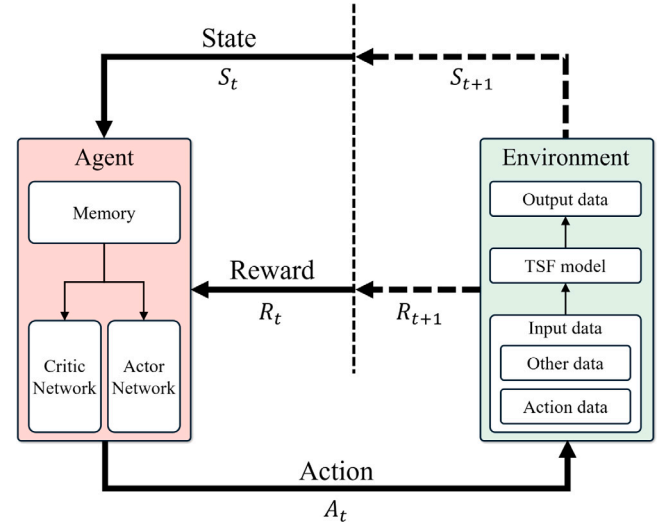


Fig. 4. The proposed TSF-PPO integrated framework for HVAC prediction and control.

formula ensures that policy updates are constrained to prevent significant deviations from the previous policy, thereby reducing unnecessary updates and ensuring learning stability. In practice, overly large policy updates can lead to performance degradation or convergence failure; thus, PPO encourages careful policy refinement within a limited range. These features are particularly useful in complex control environments with continuous state and behavior spaces, such as HVAC systems, and allow for effective learning of tradeoffs between the conflicting goals of reducing energy consumption and maintaining thermal comfort.

3.7. PPO agent training

The PPO-based control agent utilizes a GRU-based TSF model, which has demonstrated high accuracy in time series prediction, to predict the future state of the HVAC system. This predicted state serves as input to the agent to determine the control action. The control variables employed by the PPO agent include the AHU fan speed and PICV opening rate, both of which are defined within a continuous action space. The agent was trained using real operation data from a building, collected between July and September, and iteratively updates its policy based on state-action-reward-next state transitions (s_t, a_t, r_t, s_{t+1}) at each time step.

The objective of the agent's learning is to minimize energy consumption while maintaining indoor thermal comfort, measured by the PMV. To achieve this, we developed the following multi-objective reward function:

$$R_{\text{total}} = R_{\text{energy}} + R_{\text{PMV}} \quad (13)$$

Here, the energy compensation term, which promotes the minimization of energy consumption, is defined as follows:

$$R_{\text{energy}} = -E_{\text{HVAC},t}^{(\text{kWh})} \quad (14)$$

E_{HVAC} represents the normalized energy consumption estimate of the HVAC system at point t . The PMV-based comfort compensation term is calculated by applying the ANSI/ASHRAE 55-2020 standard criteria for a total of five zones and is defined at each t as follows:

$$R_{\text{PMV}} = \frac{1}{T} \sum_{t=1}^T \left(\frac{1}{Z} \sum_{i=1}^Z R_{\text{PMV},i,t} \right) \cdot \mathbb{I}_{\text{op}}(t), \quad (15)$$

$$\text{where } R_{\text{PMV},i,t} = \begin{cases} 1 & \text{if } |\text{PMV}_{i,t}| \leq 0.5 \\ 0 & \text{if } 0.5 < |\text{PMV}_{i,t}| \leq 1.5 \\ -1 & \text{if } |\text{PMV}_{i,t}| > 1.5 \end{cases}$$

where $Z = 5$ is the number of zones, T is the number of time steps, and $\mathbb{I}_{\text{op}}(t)$ is an indicator function that equals 1 during HVAC operating hours and 0 otherwise. This prevents PMV data outside of operating hours from affecting the reward. With this reward structure, the PPO agent updates its policy to maximize the cumulative reward value over time, thereby optimizing the following objective function. Fig. 4 illustrates the proposed HVAC optimal control framework, which integrates the GRU-based time series prediction model with the PPO-based control agent. The framework consists of two modules interacting and learning recursively. First, the GRU prediction model takes historical sensor data as input and predicts the future values of key variables, such as AHU power consumption and indoor temperature and humidity. These predicted values are then passed to the control agent as a state, which is utilized to determine the state of the HVAC system at the next time step. Based on this state information, the PPO control agent determines the optimal control action, which consists of continuous control variables such as the AHU fan speed and the PICV opening rate for each floor. After the control action is applied, the resulting actual energy consumption and room PMV are calculated, and a reward is computed accordingly. The reward function is specifically designed to balance the trade-off between minimizing energy consumption and maintaining comfort, and the agent uses this value to update its policy. This process is repeated at each time step, enabling the agent to gradually learn more robust and stable control policies. The GRU prediction model reflects the dynamic nature of the nonlinear time series to enhance the accuracy of future state prediction, and the PPO agent learns adaptive control strategies based on the predicted future information to simultaneously improve the efficiency and comfort of HVAC operation.

It is important to note that the predicted variables provided by the GRU model do not constitute the entire state space of the PPO agent. Instead, they serve as anticipatory features that complement real-time measurements. Specifically, the state vector used for PPO training is composed of (i) the GRU-predicted HVAC energy consumption, indoor air temperature, and relative humidity, (ii) current measured indoor environmental variables, (iii) HVAC operational states such as AHU fan speed and PICV opening rates, (iv) occupancy-related signals obtained from passive infrared (PIR) sensors, and (v) outdoor weather conditions. This design allows the agent to exploit both instantaneous observations and short-horizon forecasts, thereby mitigating the effects of unobserved disturbances and short-term uncertainties, including stochastic occupant behavior, intermittent door openings, and rapid indoor-outdoor thermal interactions.

Although the GRU forecaster predicts a limited number of target variables, the overall state space remains sufficiently informative and structured for reinforcement learning-based control. The selection of PPO is motivated not by the dimensionality of the input space, but by its suitability for continuous control under comfort constraints. PPO employs a clipped surrogate objective that ensures stable and conservative policy updates, which is particularly important in HVAC systems where abrupt control actions may degrade occupant comfort or system reliability. As a result, PPO provides a robust balance between learning stability, control smoothness, and practical applicability in real building environments.

3.8. Baseline rule-based control (RBC)

For performance comparison, the *baseline rule-based control (RBC)* strategy was adopted. This baseline represents the control logic that is *currently implemented and actively used* in the target building during the data collection period, rather than a hypothetical or simplified reference. Specifically, it reflects a typical cooling-season operational strategy implemented in the Building Energy Management System (BEMS), without any optimization or learning-based components. The baseline control follows an outdoor air enthalpy-based strategy, in which cooling operation levels are determined using outdoor air temperature

and humidity conditions. Under this strategy, control actions are executed according to predefined step-wise rules. Concretely, the AHU fan speed is regulated using discrete frequency levels, while the opening rate of PICVs is adjusted on a floor-by-floor basis to supply chilled water to FCUs. From the chilled-water-side perspective, the baseline assumes a fixed temperature difference of approximately 5 °C between the supply and return water during cooling operation. Under this assumption, the primary control objective is not to actively regulate zone-level thermal trajectories, but rather to maintain stable chilled-water circulation and system-level operational consistency. It should be noted that, in environments where indoor sensing infrastructure is unavailable, outdoor enthalpy-based control can be considered a reasonable and widely adopted strategy. However, the target building in this study is equipped with rich indoor environmental sensing, including zone-level temperature and humidity measurements as well as occupancy-related signals obtained from passive infrared (PIR) sensors. Despite the availability of this detailed indoor information, the conventional baseline does not explicitly utilize indoor environmental states in its control logic. To address this limitation, the proposed prediction-guided GRU-PPO framework explicitly leverages indoor environmental data to enable fine-grained and anticipatory control. By incorporating predicted indoor conditions into the control policy, the proposed approach demonstrates that meaningful energy savings can be achieved through indoor-state-aware operation, beyond what is attainable with outdoor-enthalpy-based rule control alone.

4. Experiment settings

4.1. Data description

The Energy data type consists of one variable, HVAC system energy consumption, which represents the energy consumed by the HVAC system per 15 min. This variable quantitatively expresses the total energy load of the system and serves as one of the key optimization targets for the reinforcement learning-based model proposed in this study. The HVAC system data type consists of two control variables: AHU fan speed and PICV opening rate. AHU fan speed refers to the rotation speed of the AHU fan, which regulates the amount of supply air and ventilation. The PICV opening rate indicates the chilled water flow control status for each floor and is designed to enable precise flow adjustment based on the cooling load. The Indoor environment data type captures the thermal environment and air quality conditions of corridors and indoor spaces. In addition to variables such as temperature, humidity, and wind speed in the corridor and indoor spaces, indoor CO_2 concentration is also measured, which is an important indicator that indirectly reflects the occupancy and ventilation status. The Occupancy data type is obtained through passive infrared (PIR) sensors installed in the corridors and rooms, which can be used to determine the occupancy of the space over time. This information is utilized for demand response-based control of the HVAC system. The Outdoor weather data type includes a variety of external environmental variables such as outdoor temperature, humidity, insolation, wind speed, rainfall, and cloud cover. These variables influence the cooling load and indoor thermal environment, which fluctuate over time, and are important factors in the control strategy of the HVAC system. The TSF model is designed to predict three key variables: HVAC system energy consumption, indoor relative humidity, and indoor air temperature. The model's input variables include all variables listed in Table 1, excluding the three target variables being predicted. The outputs generated by the trained TSF model are incorporated into the DRL state as short-horizon look-ahead features, together with the measured indoor/outdoor variables and HVAC operational signals. The actions of the DRL model consist of HVAC system control variables, namely AHU fan speed and PICV opening rate, which regulate the system's operation. These actions drive the environmental changes at the next time point, and the updated variables are then used as input

Table 1
Description of collected variables categorized by data type for HVAC energy prediction and control.

Data type	Variable name	Range	Unit
Energy	HVAC system energy consumption	0–22.88	kWh
HVAC system	AHU fan speed	0–60	Hz
	PICV opening rate	0–100	%
Indoor environment	Indoor hallway relative humidity	38.6–69.7	%
	Indoor hallway air temperature	22.6–32.5	°C
	Indoor hallway wind speed	0–0.2	m/s
	Indoor room relative humidity	41.9–65.3	%
	Indoor room air temperature	18.7–30.3	°C
	Indoor room wind speed	0–0.4	m/s
	Indoor room CO ₂ concentration	440.6–924.5	ppm
Occupancy	Passive infrared sensor in the hallway	0–900	–
	Passive infrared sensor in the room	0–900	–
Outdoor weather	Relative humidity	22.8–81.6	%
	Air temperature	20.1–40	°C
	Solar radiation	0–1241	W/m ²
	Wind speed	0–8	m/s
	Rainfall	0–55.4	mm
	Solar insolation	0–3.48	MJ/m ²
	Cloud amount	0–10	deciles
	Surface temperature	21.4–51.8	°C

to the TSF model. This structure enables the DRL model to learn an optimal control policy that balances the tradeoff between energy consumption and thermal comfort (PMV). By performing simulations based on measured operational data, data-driven control strategies applicable in situations that closely mirror real-world building environments can be derived.

4.2. Data processing

All data used in this paper are recorded at 15-minute intervals. A total of 50 data channels (variables) were collected from the IoT/BEMS infrastructure and used for model development. The data collected from the instruments and sensors contained intermittent communication errors. To address this, the linear interpolation method was applied to estimate missing values. For energy consumption and key environmental variables, missing values were compensated by using values from previous time points. Outliers were identified based on the mean value and distribution of each variable over time. After establishing an acceptable range based on the variable's characteristics, values falling outside of this range were replaced with data from a previous point in time, or with data from before and after that time. Following preprocessing, all variables were normalized using Min-Max Scaling, which adjusts each variable's value to a range between 0 and 1 while preserving the overall distribution. The normalization equation is given by:

$$x_{\text{normalized}} = \frac{x - X_{\min}}{X_{\max} - X_{\min}} \quad (16)$$

where x represents the original data, $X_{\max}$ and $X_{\min}$ denote the maximum and minimum values of the variable in question, respectively, and $x_{\text{normalized}}$ is the normalized result. Min-Max normalization enhances model stability during training by unifying the scale between variables. Although it is somewhat sensitive to outliers, this method is well-suited for datasets with continuous variables, such as those in HVAC systems (Patro and Sahu, 2015). In this study, normalized data were used throughout the training process of both the TSF and reinforcement learning models.

4.3. Hyperparameter tuning

An exhaustive grid search was conducted to fine-tune the hyperparameters of both LSTM and GRU models. The search space included sequence length {1, 4, 6, 12, 24, 48, 96}, batch size {12, 32, 64, 128, 256}, and number of training epochs {10, 20, 30, 40, 50}. For fair comparison,

Table 2
Grid-search hyperparameter candidates for TSF models (LSTM and GRU).

Hyperparameter	Candidate set
Sequence length (k)	1, 4, 6, 12, 24, 48, 96
Batch size	12, 32, 64, 128, 256
Epochs	10, 20, 30, 40, 50

Table 3
Best hyperparameters selected by grid search based on validation loss.

Model	Sequence length	Batch size	Epochs	Validation loss
LSTM	96	32	50	0.000637
GRU	96	32	50	0.000577

both models were trained using identical datasets, loss functions, and optimization settings. The optimal configuration was selected based on validation loss, ensuring robust and stable forecasting performance for subsequent reinforcement learning control. For fair comparison, the same sequence length and batch size were also adopted for the LSTM model. Tables 2 and 3 summarize the grid-search procedure used to fine-tune the TSF models. We explored the sequence length k , batch size, and the number of epochs as key training hyperparameters, and selected the best configuration based on the lowest validation loss.

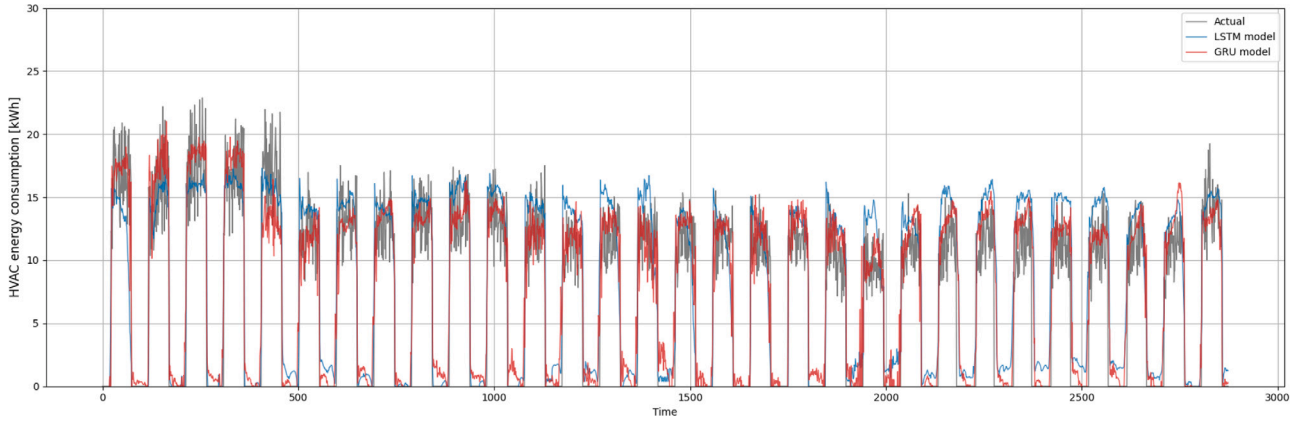
5. Results and discussion

In this study, a reinforcement learning agent was trained to optimize the control of an HVAC system using the PPO algorithm. The experimental results are divided into two main parts: (1) performance evaluation of the TSF model and (2) performance analysis of the reinforcement learning-based HVAC control system trained using the TSF model. The prediction model was trained using the TensorFlow framework in the Python environment. Mean Squared Error (MSE) was employed as the loss function, and the network weights were updated using the Adam Optimizer. The TSF models (LSTM and GRU) were trained in TensorFlow using the Adam optimizer and mean squared error (MSE) loss. Hyperparameters were selected via grid search (Tables 2–3), and the best configuration was sequence length $k = 96$, batch size 32, and 50 epochs, based on the lowest validation loss. Model checkpointing was used to save the best model on the validation set. Model checkpointing was configured to automatically save the best-performing model based on the validation dataset. The experimental

Table 4

Comparison of LSTM and GRU performance for each target variable.

Target	RMSE		R^2		SMAPE	
	LSTM	GRU	LSTM	GRU	LSTM	GRU
Energy consumption	3.1042	2.5119	0.7819	0.8572	–	–
Indoor 1F Room 1 relative humidity	2.1222	1.1016	0.8530	0.9604	3.0195	1.3818
Indoor 1F Room 1 air temperature	0.6471	0.8636	0.8744	0.7764	2.2298	3.3154
Indoor 1F Room 2 relative humidity	2.1680	2.2832	0.8584	0.8430	3.3951	3.5492
Indoor 1F Room 2 air temperature	0.6413	0.4417	0.9158	0.9601	2.1708	1.3927
Indoor 2F Room 1 relative humidity	2.2118	2.0147	0.7789	0.8165	3.3169	3.4550
Indoor 2F Room 1 air temperature	0.6818	0.5462	0.8486	0.9028	2.4157	1.8292
Indoor 3F Room 1 relative humidity	1.5457	1.3964	0.9185	0.9335	2.3812	2.1042
Indoor 3F Room 1 air temperature	0.9657	0.6078	0.8156	0.9270	3.3808	2.0719
Indoor 3F Room 2 relative humidity	2.1223	1.3884	0.8732	0.9458	3.3171	1.9906
Indoor 3F Room 2 air temperature	1.3464	0.6049	0.6365	0.9266	5.1283	2.1338

**Fig. 5.** Comparison of predicted and actual HVAC system energy consumption (LSTM vs. GRU).

setup for training the TSF and DRL models was based on an Intel i5-13600KF CPU, Nvidia GeForce RTX 4060 Ti GPU, and 32 GB of RAM.

5.1. TSF model evaluation metrics

This study evaluates the robustness of the proposed prediction model using three metrics: Root Mean Square Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Coefficient of Determination (R^2). These metrics are commonly used to evaluate TSF models and are defined by the following formulas:

$$RMSE(A_t, P_t) = \sqrt{\frac{1}{n} \sum_{t=1}^{n-1} (A_t - P_t)^2} \quad (17)$$

$$SMAPE(A_t, P_t) = \frac{100\%}{n} \sum_{t=1}^n \frac{|P_t - A_t|}{\left(\frac{|A_t| + |P_t|}{2}\right)} \quad (18)$$

$$R^2(A_t, P_t) = 1 - \frac{\sum_{t=1}^n (A_t - P_t)^2}{\sum_{t=1}^n (A_t - \bar{A})^2} \quad (19)$$

where A_t represents the actual value at time t , P_t is the predicted value, n denotes the total number of time points, and $\bar{A}$ indicates the average of the actual values over the entire time period. RMSE quantifies the absolute error between the forecasted and actual values, with smaller values indicating higher forecast accuracy. SMAPE, on the other hand, calculates the error between the predicted and actual values, normalized by the average of the two values, thus addressing the limitation of traditional metrics, which experience an increase in error when the actual value is zero. SMAPE values range between 0% and 200%, with 0% indicating optimal forecast performance. R^2 is an explanatory power metric that indicates how well the predicted value accounts for the variance of the true value, with values approaching 1 suggesting a higher model performance. SMAPE, as a ratio-based metric, has the drawback of generating large errors when the predicted

or actual value approaches zero. Therefore, for variables that may contain zero values, such as energy consumption, R^2 was used as the main evaluation metric. Conversely, for variables that consistently maintain positive values, such as temperature and humidity, SMAPE was applied to assess the prediction performance. Furthermore, the performance of the reinforcement learning-based control model was evaluated based on the cumulative energy consumption savings over the same period, compared to the conventional rule-based operation method. The effectiveness of the control policies for different operational purposes was also verified by comparing the performance of the model that prioritizes energy efficiency with that of the model that prioritizes occupant thermal comfort (PMV), achieved by adjusting the coefficients of the reward function.

5.2. TSF model

In this section, the performance of the proposed Time Series Forecasting (TSF) model is quantitatively evaluated, and a comparative analysis between the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models in predicting energy consumption and indoor environmental variables of HVAC systems is performed. RMSE (Root Mean Squared Error), R^2 (Coefficient of Determination), and SMAPE (Symmetric Mean Absolute Percentage Error) were used as evaluation metrics, and the prediction accuracy of the models was verified using temperature and humidity data for each indoor space.

Table 4 shows the results of the performance comparison between the proposed time series prediction models, LSTM and GRU, based on various evaluation metrics (RMSE, R^2 , and SMAPE). Energy consumption and indoor environmental variables (temperature and humidity) were set as the prediction targets, and the prediction accuracy of the two models was compared for each variable. The GRU model demonstrated superior performance in predicting energy consumption, achieving an RMSE of 2.5119 and R^2 of 0.8572, compared to the

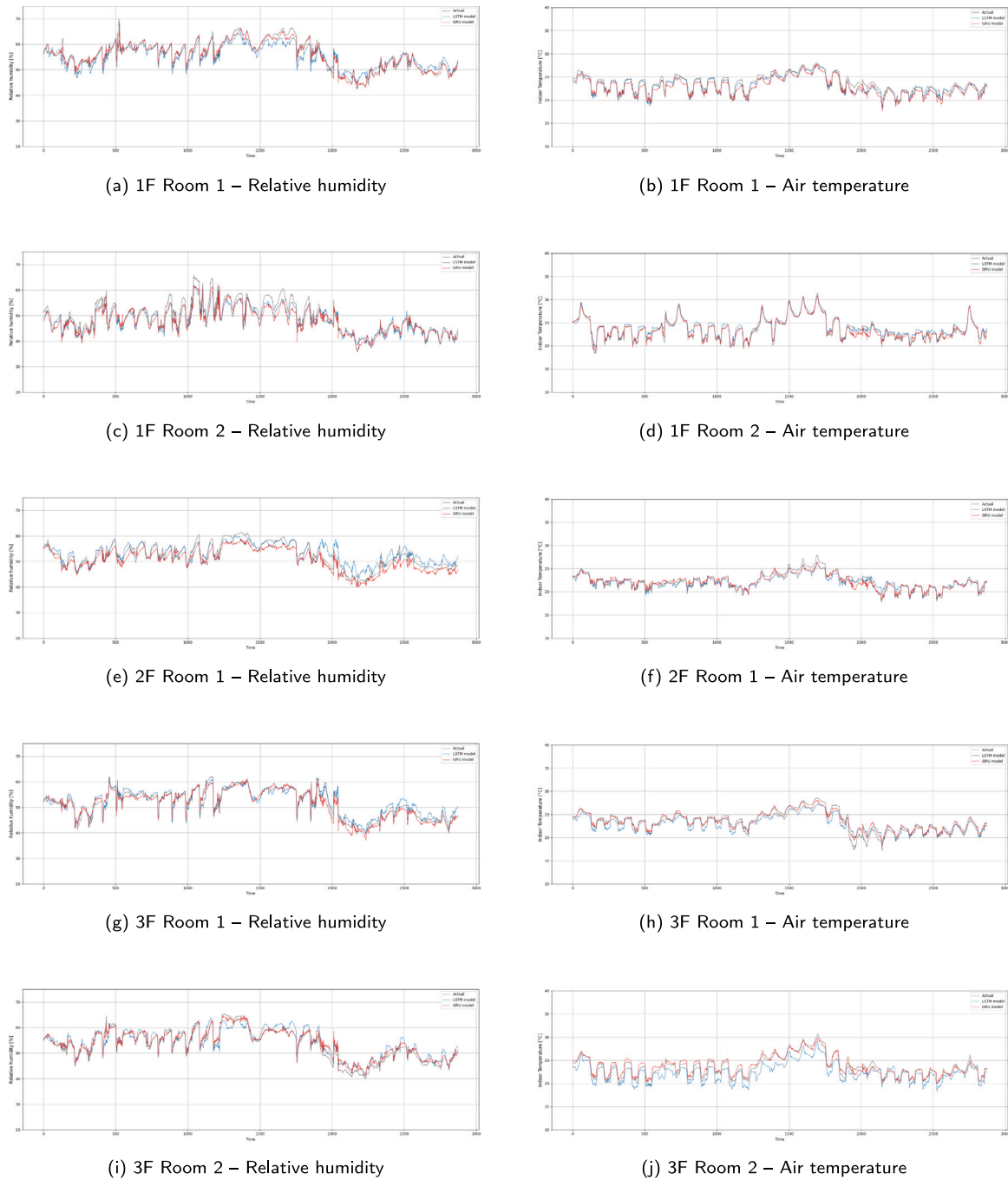


Fig. 6. Comparison of predicted and actual indoor environment (LSTM vs. GRU).

LSTM model, which recorded an RMSE of 3.1042 and R^2 of 0.7819. Although SMAPE was not calculated for this variable, the GRU model outperformed the LSTM model on absolute error-based metrics. For indoor humidity prediction, GRU demonstrated lower RMSE and SMAPE values and higher R^2 values than LSTM in most rooms. Notably, for humidity prediction in Room 1 on the first floor, the GRU model significantly outperformed the LSTM model, achieving an RMSE of 1.1016, R^2 of 0.9604, and SMAPE of 1.3818 compared to those of LSTM (RMSE 2.1222, R^2 0.853, and SMAPE 3.0195). These results indicate that GRU is more effective at predicting variables with relatively slow dynamic changes, such as humidity. For temperature prediction, the GRU model also demonstrated superior performance in most rooms, particularly in Room 2 on the third floor, where it achieved an RMSE of 0.6049, R^2

of 0.9266, and SMAPE of 2.1338. In contrast, the LSTM model showed relatively poor performance with an RMSE of 1.3464, R^2 of 0.6365, and SMAPE of 5.1283. However, for temperature prediction in Room 1 on the first floor, the LSTM model outperformed the GRU model, yielding lower RMSE and SMAPE values. Overall, GRU consistently exhibited lower prediction errors and higher explanatory power than LSTM across most variables, with better overall stability and precision, especially for indoor humidity and temperature predictions. Therefore, GRU was selected as the final TSF model and was subsequently utilized as input for the reinforcement learning-based control model, enhancing both the effectiveness and practicality of prediction-based HVAC system control (see Figs. 6 and 8).

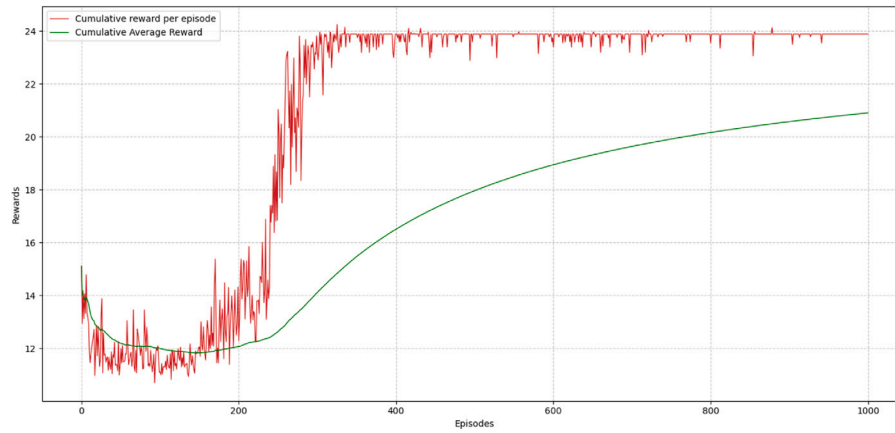


Fig. 7. Learning curve of the prediction-guided GRU-PPO agent, showing per-episode cumulative reward and running average reward.

Table 5

Monthly comparison of energy consumption and thermal comfort (PMV) under different control strategies.

Month	Energy consumption[kWh]			PMV			Relative Energy Variation [%]	
	RBC	PPO	GRU-PPO	RBC	PPO	GRU-PPO	PPO	GRU-PPO
Jul.	21,477.18	21,258.02	20,021.03	-0.78	-0.75	+0.37	-1.02	-6.78
Aug.	26,204.77	26,745.82	23,099.50	+0.15	+0.12	+0.31	+2.06	-11.85
Sep.	19,384.67	18,724.13	17,713.71	-0.41	-0.38	+0.37	-3.04	-8.62

Based on the above analysis results, the GRU-based TSF model outperformed the LSTM-based model in predicting the HVAC system energy consumption and indoor temperature and humidity, demonstrating an overall lower prediction error and higher explanatory power. Specifically, the time series graphs presented in Figs. 4–5 illustrate that the GRU model more closely follows the trend of the actual measurements, and the learning results are more sophisticated in terms of prediction stability and response sensitivity. Based on these results, the GRU model was selected as the final TSF in this study, and its prediction results were subsequently input as the state of the reinforcement learning-based controller (PPO) to simulate the optimal energy control of the HVAC system. This enabled the implementation of an integrated energy management strategy that simultaneously achieves prediction accuracy and control efficiency.

5.3. DRL

The DRL environment was implemented in Python using the OpenAI Gym toolkit (Brockman et al., 2016) and is driven by measured building data rather than a physics-based simulator. A GRU-based TSF model was used to provide short-horizon forecasts of indoor temperature, relative humidity, and HVAC energy consumption, which were appended to the PPO state as look-ahead features. The agent then outputs continuous actions (AHU fan speed and PICV opening rate), and the resulting energy consumption and PMV are computed to construct the reward. For comparison, a PPO agent trained without forecast features was evaluated under the same environment and reward design. This setup enables a controlled assessment of prediction-guided versus reactive control under identical operating conditions.

Fig. 7 illustrates the reward evolution during the training of the prediction-guided GRU-PPO agent. The horizontal axis represents the number of training episodes (0–1000), while the vertical axis denotes the cumulative reward obtained within each episode. The red curve corresponds to the per-episode cumulative reward, whereas the green curve shows the running (cumulative) average reward, which provides a smoothed indication of learning progress. In the early training phase, the agent exhibits low reward values with high variability, reflecting exploratory behavior and policy instability. After approximately 300 episodes, a clear upward trend is observed, and both the episode-wise

reward and its running average begin to stabilize. The convergence of the running average reward indicates that the GRU-PPO agent progressively learns more effective control policies. Overall, these results indicate stable learning behavior of the prediction-guided GRU-PPO framework and suggest its potential to derive effective HVAC control strategies.

Table 5 summarizes the monthly comparison of HVAC energy consumption and thermal comfort (PMV) under three control strategies: baseline rule-based control (RBC), PPO without forecast input, and prediction-guided PPO (GRU-PPO), during the peak cooling months from July to September. Compared with the baseline RBC, the PPO controller without forecast input achieved limited and less consistent energy savings across months, reducing energy consumption by 1.02% in July and 3.04% in September, while slightly increasing energy use by 2.06% in August. In contrast, the proposed prediction-guided GRU-PPO consistently outperformed both RBC and PPO across all months, achieving energy reductions of 6.78% in July, 11.85% in August, and 8.62% in September. These results indicate that incorporating short-horizon forecasts enhances energy-saving performance, particularly during high cooling-load periods.

In terms of thermal comfort, RBC produced PMV values biased toward cooler conditions (e.g., -0.78 in July), reflecting its lack of zone-level indoor feedback. The PPO controller without forecast input exhibited only marginal differences from RBC, yielding similar monthly-average PMV values (e.g., -0.75 in July and -0.38 in September). In contrast, GRU-PPO shifted PMV toward the comfort band and maintained a more stable profile across months (monthly-average PMV: +0.31 to +0.37). These results indicate that prediction-guided control enables proactive adjustment of HVAC operation, effectively balancing energy efficiency and occupant comfort. This indicates that prediction-guided control enables proactive adjustment of HVAC operation, effectively balancing energy efficiency and occupant comfort. Overall, the results confirm that the proposed GRU-PPO framework achieves superior and more stable performance than both conventional rule-based control and reactive PPO control, highlighting the benefit of prediction-guided state design in real-building HVAC optimization.

For each month, we selected a representative “best energy-saving day” defined by the maximum daily energy reduction of GRU-PPO relative to RBC. Fig. 9 compares the AHU supply fan speed trajectories under RBC, PPO without forecast input, and prediction-guided

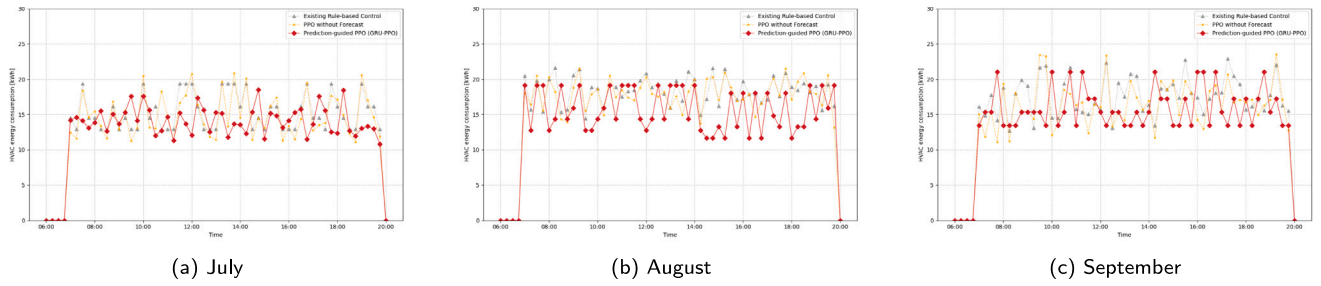


Fig. 8. HVAC energy-consumption profiles on the best energy-saving day (Jul–Sep): Existing Rule-based Control vs. PPO without Forecast vs. Prediction-guided PPO (GRU-PPO).

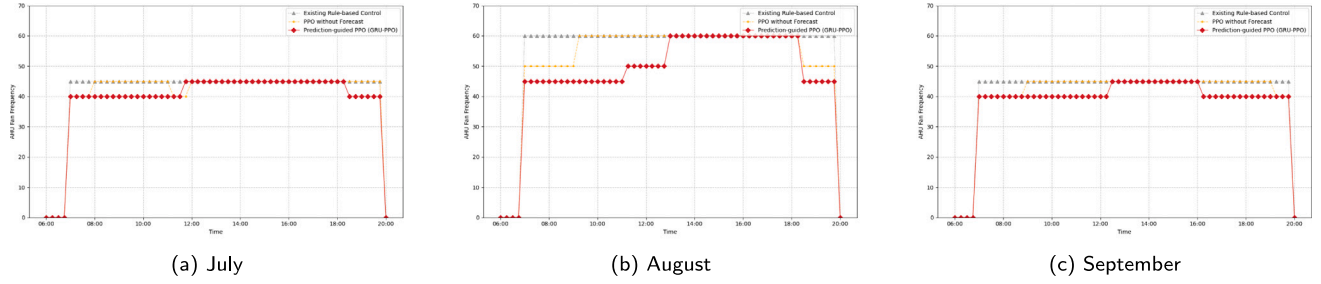


Fig. 9. AHU supply fan frequency on the best energy-saving day (Jul–Sep): Existing Rule-based Control vs. PPO without Forecast vs. Prediction-guided PPO (GRU-PPO).

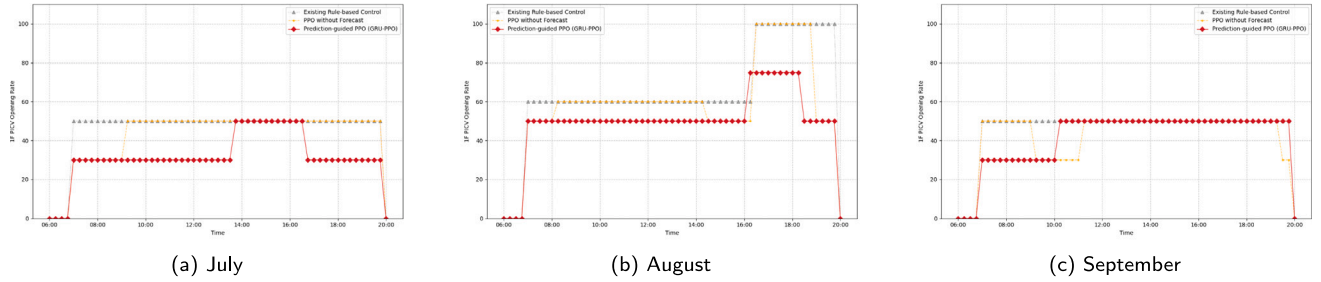


Fig. 10. PICV opening-rate profiles on the best energy-saving day (Jul–Sep): Existing Rule-based Control vs. PPO without Forecast vs. Prediction-guided PPO (GRU-PPO).

PPO (GRU-PPO) on representative high energy-saving days. The RBC strategy maintained relatively high and nearly constant fan speeds (approximately 45–60 Hz) throughout the operating period, reflecting its rule-based and non-adaptive nature. The PPO controller without forecast input exhibited partial adaptability by adjusting fan speeds in response to observed load variations; however, its responses were often reactive and less anticipatory. In contrast, the GRU-PPO controller demonstrated more proactive and smoother fan speed modulation. For example, in July, GRU-PPO reduced the fan speed to approximately 40 Hz during periods of reduced cooling demand, while maintaining sufficient airflow to preserve indoor comfort. Similar trends were observed in August and September, where GRU-PPO consistently operated at lower fan speeds during off-peak load periods, particularly in the morning and late afternoon. These results indicate that incorporating short-horizon forecasts enables the control agent to anticipate load changes and minimize unnecessary fan operation, thereby improving HVAC energy efficiency.

Fig. 10 presents a comparison of PICV opening-rate control under RBC, PPO without forecast input, and GRU-PPO. Under the RBC strategy, PICV openings remained largely fixed within a high range of 50%–75% during cooling operation, regardless of temporal load fluctuations. The PPO controller without forecast input showed improved adaptability compared to RBC, but its adjustments were primarily driven by instantaneous observations. In contrast, the GRU-PPO controller exhibited finer and more anticipatory regulation of PICV opening

rates, dynamically adjusting within a wider range of approximately 30%–75% in response to predicted cooling demand. Notably, during afternoon periods in August and September, GRU-PPO actively reduced PICV openings under declining loads, preventing excessive chilled-water flow while maintaining thermal comfort. This prediction-guided modulation effectively reduces unnecessary pump operation and cooling energy consumption, demonstrating the advantage of forecast-informed control over purely reactive strategies.

Fig. 11 compares the floor-level PMV trajectories under RBC, PPO without forecast input, and prediction-guided PPO (GRU-PPO). The PMV profiles obtained using PPO without forecast input closely resemble those of the conventional rule-based control. In both cases, PMV values tend to deviate toward cooler conditions during peak cooling periods and exhibit limited responsiveness to short-term thermal disturbances. This similarity indicates that, without predictive state information, the PPO agent behaves in a largely reactive manner and fails to substantially improve indoor comfort beyond what is achieved by the baseline RBC strategy. Although minor fluctuations are observed due to stochastic policy exploration, the overall PMV trends remain comparable to those of RBC across all floors. In contrast, GRU-PPO keeps PMV closer to the ANSI/ASHRAE 55 comfort band ($[-0.5, +0.5]$) across floors on representative high-load summer days, showing reduced deviations during periods of rapid load variation. By incorporating short-horizon forecasts of indoor environmental states

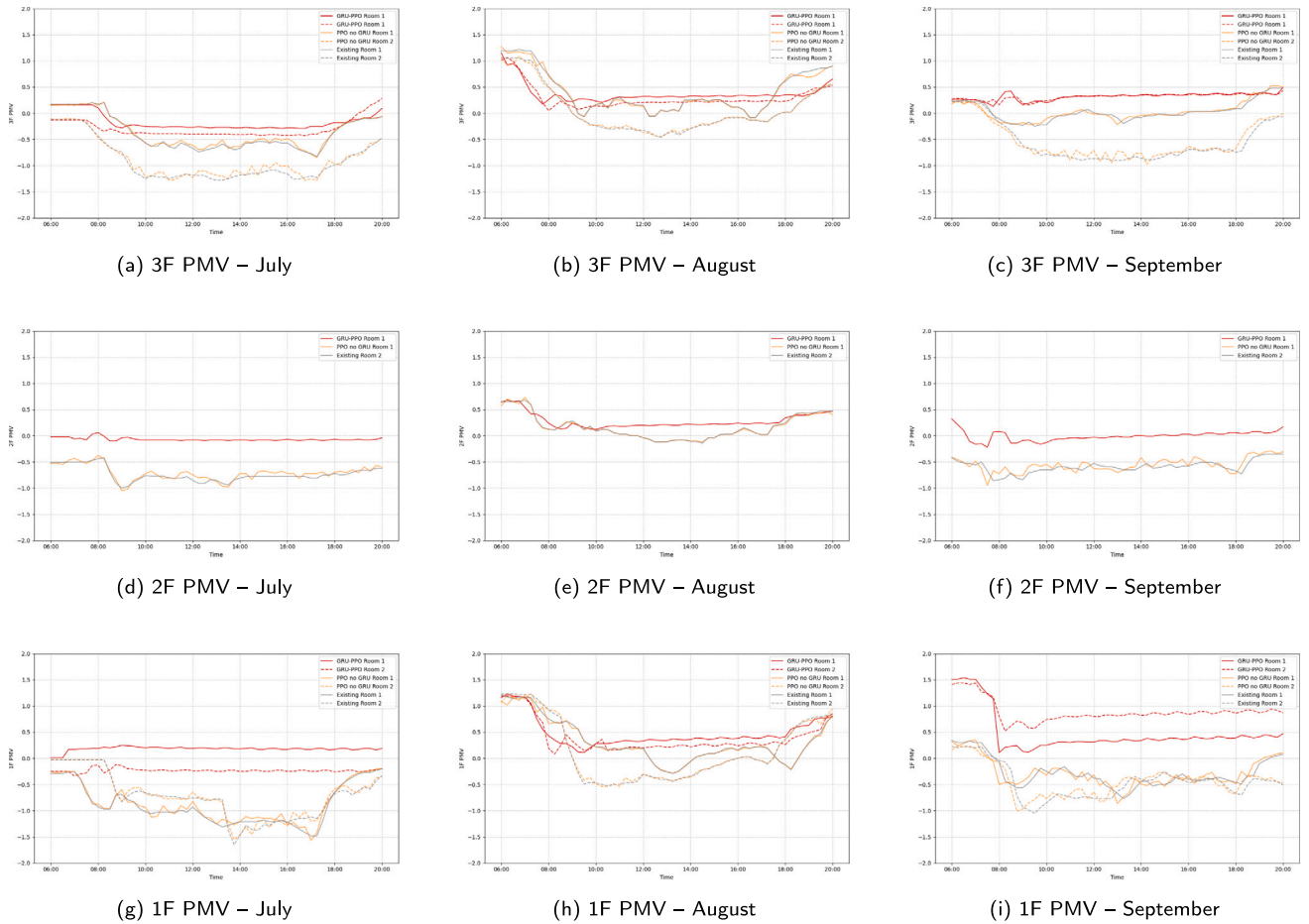


Fig. 11. Floor-level PMV trajectories on the best energy-saving day (Jul-Sep): Existing Rule-based Control vs. PPO without Forecast vs. Prediction-guided PPO (GRU-PPO).

Table 6

Computational energy consumption and carbon emissions during model development.

Stage	Electricity consumption (kWh)	Estimated emissions (kgCO ₂ eq)	Share (%)
Data preprocessing	9.0×10^{-6}	3.87×10^{-6}	< 0.01
TSF training	0.0987	0.0425	27.1
DRL training	0.2666	0.1148	72.9
Total	0.3653	0.1573	100.0

into the control state, GRU-PPO anticipates thermal dynamics and proactively adjusts control actions. This prediction-guided behavior results in significantly improved PMV stability and comfort preservation, particularly during periods of rapid load variation in summer operation.

Overall, the comparative analysis of control actions and comfort outcomes highlights a clear distinction between reactive and prediction-guided control paradigms. While the PPO controller without forecast input exhibits limited improvements over the baseline RBC, its behavior remains largely reactive and closely follows conventional operation patterns. In contrast, the proposed GRU-PPO framework consistently demonstrates anticipatory control behavior across fan speed modulation, PICV regulation, and PMV stabilization. By leveraging short-horizon forecasts of indoor environmental states, GRU-PPO enables coordinated and proactive adjustments of HVAC operation, leading to reduced energy consumption and improved thermal comfort stability during periods of rapid load variation.

In addition to operational energy savings, the computational energy consumption and associated carbon emissions incurred during model training were quantitatively assessed using the CodeCarbon framework. Table 6 summarizes the electricity usage and estimated

carbon emissions for each major training stage. The data preprocessing stage incurred negligible computational cost, consuming approximately 9×10^{-6} kWh and emitting 3.87×10^{-6} kgCO₂eq. Training of the GRU-based time-series forecasting model consumed 0.0987 kWh of electricity, corresponding to an estimated 0.0425 kgCO₂eq. The PPO-based reinforcement learning training stage required 0.2666 kWh, resulting in 0.1148 kgCO₂eq. Overall, the total electricity consumption for model development amounted to approximately 0.365 kWh, with a cumulative carbon emission of 0.157 kgCO₂eq. Based on the experimental results, the proposed GRU-PPO framework achieved an average HVAC energy reduction of approximately 9.1%, corresponding to a total energy saving of 6103 kWh over the evaluation period (from an original consumption of 67,066.62 kWh). Using a conservative grid emission factor of 0.424 kgCO₂/kWh, this energy reduction translates into an operational carbon emission reduction of approximately 2588 kgCO₂. In contrast, the cumulative carbon emissions associated with model training amounted to approximately 0.157 kgCO₂eq. Consequently, the carbon emissions incurred during model development are offset after only a very short period of operational deployment. This result indicates that the proposed prediction-guided DRL framework

delivers substantial net carbon benefits, even when accounting for its computational training cost.

6. Conclusion

This study proposed a prediction-guided deep reinforcement learning framework for optimizing HVAC operation in multi-zone smart buildings equipped with pressure-independent control valves (PICVs). By integrating a GRU-based time-series forecasting (TSF) model with a Proximal Policy Optimization (PPO) agent, short-horizon predictions of indoor environmental conditions and HVAC energy consumption were incorporated into the control state, enabling anticipatory and data-driven decision making. Experimental results using real building operational data from July to September demonstrated that the proposed GRU-PPO framework consistently outperformed both conventional rule-based control (RBC) and reactive PPO control without forecast input. The prediction-guided controller achieved up to 11.85% monthly energy savings and an average reduction of approximately 9.1% across the evaluation period, while maintaining indoor thermal comfort within the ANSI/ASHRAE 55 recommended PMV range of $[-0.5, +0.5]$. Analysis of control behaviors revealed that GRU-PPO enabled smoother and more proactive modulation of AHU fan speeds and PICV opening rates, effectively suppressing unnecessary system operation during low-load periods while preserving occupant comfort. Beyond performance improvement, this study demonstrates that the benefit of prediction-guided control extends beyond algorithm substitution. By injecting forecast information into the control state, the GRU forecaster fundamentally alters the control paradigm, allowing the reinforcement learning agent to anticipate thermal dynamics rather than responding solely to instantaneous measurements. Comparative experiments showed that a PPO controller without forecast input behaved similarly to the baseline RBC, particularly in terms of PMV trajectories, underscoring the critical role of forecast-informed state design in real building HVAC optimization.

From a sustainability perspective, the computational overhead associated with model training was modest. The total electricity consumption for training the TSF and DRL models was approximately 0.37 kWh, corresponding to an estimated 0.16 kgCO₂eq. When compared with the cumulative HVAC energy savings achieved during operation, this training-related carbon footprint can be offset within a short operational period. This result indicates that the proposed framework is environmentally justified from a lifecycle carbon perspective.

The proposed framework is also well aligned with emerging building-level distributed energy systems. In particular, the integration of prediction-guided HVAC control with on-site low-carbon power sources, such as fuel cells, energy storage systems (ESS), and photovoltaics (PV), represents a promising direction for further reducing operational emissions. Fuel-cell-based power generation offers stable, dispatchable electricity and waste-heat recovery potential, making it especially suitable for coupling with HVAC systems. By coordinating HVAC demand through predictive DRL control with fuel cell operation schedules, additional system-level efficiency gains and carbon reductions can be achieved beyond HVAC optimization alone. Despite these promising results, several limitations should be acknowledged. First, the control policies were evaluated in a physics-engine-free, data-driven simulation environment driven by measured operational data, rather than being deployed in real-time closed-loop control. While this approach closely reflects real building behavior, on-site implementation may introduce additional uncertainties related to sensor noise, communication delays, and actuator constraints. Second, the GRU-based forecaster predicts a limited set of variables over a short horizon; extending the prediction scope or incorporating probabilistic forecasts could further enhance robustness under highly stochastic conditions. Third, this study focused on cooling operation in a single building within a multi-building complex; extension to heating seasons, mixed-mode operation, and other building typologies remains an important

area for future research. Future work will focus on real-time deployment of the proposed framework and its extension toward integrated energy management. In particular, joint optimization of HVAC control with fuel cell generation, ESS dispatch, and PV utilization will be investigated under carbon-aware and price-responsive reward formulations. Such an extension would enable holistic demand-side management and support the realization of low-carbon, self-sustaining smart buildings and district-scale energy systems. Overall, the proposed GRU-PPO framework provides a practical, scalable, and environmentally meaningful solution for predictive HVAC control, demonstrating that forecast-informed reinforcement learning can simultaneously achieve energy efficiency, thermal comfort, and carbon reduction in real-world smart building applications.

CRedit authorship contribution statement

Younghyun Baek: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Guwon Yoon:** Writing – review & editing, Investigation, Formal analysis. **Seunghwan Kim:** Validation, Software, Methodology. **Byeongkwan Kang:** Methodology, Investigation, Conceptualization. **Sanghoon Lee:** Visualization, Investigation, Data curation. **Myeong-in Choi:** Writing – review & editing, Software, Formal analysis. **Sehyun Park:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-RS-2024-00436248, 33.4), and this work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant, funded by the Korean government (MOTIE) (RS-2024-00398346, ESS Big Data-Based O&M and Asset Management Technical Manpower Training), and this research was supported by the Chung-Ang University Graduate Research Scholarship in 2024.

Data availability

Data will be made available on request.

References

- Akter, R., Shirkoobi, M.G., Wang, J., Mérida, W., 2025. An efficient hybrid deep neural network model for multi-horizon forecasting of power loads in academic buildings. *Energy Build.* 329, 115217. <http://dx.doi.org/10.1016/j.enbuild.2024.115217>.
- Almalq, A., Zhang, J.J., 2019. Evolutionary deep learning-based energy consumption prediction for buildings. *IEEE Access* 7, 1520–1531. <http://dx.doi.org/10.1109/ACCESS.2018.2887023>.
- Amber, K.P., Ahmad, R., Aslam, M.W., Kousar, A., Usman, M., Khan, M.S., 2018. Intelligent techniques for forecasting electricity consumption of buildings. *Energy* 157, 886–893. <http://dx.doi.org/10.1016/j.energy.2018.05.155>.
- ASHRAE, 0000. ANSI/ASHRAE Addendum j to ANSI/ASHRAE Standard 55-2020.
- Baldi, S., Korkas, C.D., Lv, M., Kosmatopoulos, E.B., 2018. Automating occupant-building interaction via smart zoning of thermostatic loads: A switched self-tuning approach. *Appl. Energy* 231, 1246–1258. <http://dx.doi.org/10.1016/j.apenergy.2018.09.188>.
- Biemann, M., Scheller, F., Liu, X., Huang, L., 2021. Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. *Appl. Energy* 298, 117164. <http://dx.doi.org/10.1016/j.apenergy.2021.117164>.

- Bird, M., Daveau, C., O'Dwyer, E., Acha, S., Shah, N., 2022. Real-world implementation and cost of a cloud-based MPC retrofit for HVAC control systems in commercial buildings. *Energy Build.* 270, 112269. <http://dx.doi.org/10.1016/j.enbuild.2022.112269>.
- Blum, D., Wang, Z., Weyandt, C., Kim, D., Wetter, M., Hong, T., Piette, M.A., 2022. Field demonstration and implementation analysis of model predictive control in an office HVAC system. *Appl. Energy* 318, 119104. <http://dx.doi.org/10.1016/j.apenergy.2022.119104>.
- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W., 2016. OpenAI Gym. <http://dx.doi.org/10.48550/arXiv.1606.01540>, arXiv:1606.01540.
- Chaturvedi, S., Rajasekar, E., Natarajan, S., McCullen, N., 2022. A comparative assessment of SARIMA, LSTM RNN and Fb Prophet models to forecast total and peak monthly energy demand for India. *Energy Policy* 168, 113097. <http://dx.doi.org/10.1016/j.enpol.2022.113097>.
- Chen, Y., Norford, L.K., Samuelson, H.W., Malkawi, A., 2018. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy Build.* 169, 195–205. <http://dx.doi.org/10.1016/j.enbuild.2018.03.051>.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. <http://dx.doi.org/10.48550/arXiv.1406.1078>, arXiv:1406.1078.
- Choi, M.-i., Kang, B., Lee, S., Park, S., Seon Beck, J., Hyeon Lee, S., Park, S., 2024. Empirical study on optimization methods of building energy operation for the sustainability of buildings with integrated renewable energy. *Energy Build.* 305, 113908. <http://dx.doi.org/10.1016/j.enbuild.2024.113908>.
- Chou, J.-S., Tran, D.-S., 2018. Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. *Energy* 165, 709–726. <http://dx.doi.org/10.1016/j.energy.2018.09.144>.
- Dal Cin, E., Carraro, G., Volpato, G., Lazzaretto, A., Danieli, P., 2022. A multi-criteria approach to optimize the design-operation of Energy Communities considering economic-environmental objectives and demand side management. *Energy Convers. Manage.* 263, 115677. <http://dx.doi.org/10.1016/j.enconman.2022.115677>.
- Elmaz, F., Eyckerman, R., Casteels, W., Latré, S., Hellinckx, P., 2021. CNN-LSTM architecture for predictive indoor temperature modeling. *Build. Environ.* 206, 108327. <http://dx.doi.org/10.1016/j.buildenv.2021.108327>.
- Fang, Z., Crimier, N., Scanu, L., Midelet, A., Alyafi, A., Delinchant, B., 2021. Multi-zone indoor temperature prediction with LSTM-based sequence to sequence model. *Energy Build.* 245, 111053. <http://dx.doi.org/10.1016/j.enbuild.2021.111053>.
- Fu, Q., Li, Z., Ding, Z., Chen, J., Luo, J., Wang, Y., Lu, Y., 2023. ED-DQN: An event-driven deep reinforcement learning control method for multi-zone residential buildings. *Build. Environ.* 242, 110546. <http://dx.doi.org/10.1016/j.buildenv.2023.110546>.
- Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., Levine, S., 2019. Soft actor-critic algorithms and applications. <http://dx.doi.org/10.48550/arXiv.1812.05905>, arXiv:1812.05905.
- Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* 06 (02), 107–116. <http://dx.doi.org/10.1142/S0218488598000094>.
- Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. <http://dx.doi.org/10.1162/neco.1997.9.8.1735>.
- IEA, 2024. World Energy Outlook 2024 – Analysis. <https://www.iea.org/reports/world-energy-outlook-2024>.
- Jang, J., Han, J., Leigh, S.-B., 2022. Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks. *Energy Build.* 255, 111647. <http://dx.doi.org/10.1016/j.enbuild.2021.111647>.
- Li, Y., Song, L., Zhang, S., Kraus, L., Adcox, T., Willardson, R., Komandur, A., Lu, N., 2023. A TCN-based hybrid forecasting framework for hours-ahead utility-scale PV forecasting. *IEEE Trans. Smart Grid* 14 (5), 4073–4085. <http://dx.doi.org/10.1109/TSG.2023.3236992>.
- Minassian, R., Mihăiță, A.-S., Shirazi, A., 2025. Optimizing indoor environmental prediction in smart buildings: A comparative analysis of deep learning models. *Energy Build.* 327, 115086. <http://dx.doi.org/10.1016/j.enbuild.2024.115086>.
- Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning. In: *Proceedings of the 33rd International Conference on Machine Learning*. PMLR, pp. 1928–1937.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning. <http://dx.doi.org/10.48550/arXiv.1312.5602>, arXiv:1312.5602.
- Morteza, A., Yahyaiean, A.A., Mirzaei-bonehkhater, M., Sadeghi, S., Mohaimeni, A., Taheri, S., 2023. Deep learning hyperparameter optimization: Application to electricity and heat demand prediction for buildings. *Energy Build.* 289, 113036. <http://dx.doi.org/10.1016/j.enbuild.2023.113036>.
- Naldi, C., Dongellini, M., Morini, G.L., Rossi di Schio, E., 2023. The adoption of pressure independent control valves (PICVs) for the simultaneous optimization of energy consumption and comfort in buildings. *Energy Build.* 287, 112969. <http://dx.doi.org/10.1016/j.enbuild.2023.112969>.
- Nguyen, A.T., Pham, D.H., Oo, B.L., Santamouris, M., Ahn, Y., Lim, B.T.H., 2024. Modelling building HVAC control strategies using a deep reinforcement learning approach. *Energy Build.* 310, 114065. <http://dx.doi.org/10.1016/j.enbuild.2024.114065>.
- Ocampo Batlle, E.A., Escobar Palacio, J.C., Silva Lora, E.E., Martínez Reyes, A.M., Melian Moreno, M., Morejón, M.B., 2020. A methodology to estimate baseline energy use and quantify savings in electrical energy consumption in higher education institution buildings: Case study, Federal University of Itajubá (UNIFEI). *J. Clean. Prod.* 244, 118551. <http://dx.doi.org/10.1016/j.jclepro.2019.118551>.
- Pagare, H., Mishra, K., Sharma, K., Kotecha, K., Kulkarni, A., 2025. Reducing energy consumption in air conditioning systems, a fuzzy logic-based optimization approach. *Energy Rep.* 14, 3014–3032. <http://dx.doi.org/10.1016/j.egyr.2025.09.021>.
- Patro, S.G.K., Sahu, K.K., 2015. Normalization: A preprocessing stage. <http://dx.doi.org/10.48550/arXiv.1503.06462>, arXiv:1503.06462.
- Peng, Y., Shen, H., Tang, X., Zhang, S., Zhao, J., Liu, Y., Nie, Y., 2023. Energy consumption optimization for heating, ventilation and air conditioning systems based on deep reinforcement learning. *IEEE Access* 11, 88265–88277. <http://dx.doi.org/10.1109/ACCESS.2023.3305683>.
- Qays, M.O., Ahmad, I., Habibi, D., Moses, P., 2025. Long-term techno-economic analysis considering system strength and reliability shortfalls of electric vehicle-to-grid-systems installations integrated with renewable energy generators using hybrid GRU-classical optimization method. *Sustain. Energy Technol. Assessments* 76, 104294. <http://dx.doi.org/10.1016/j.seta.2025.104294>.
- Razzano, G., Brandi, S., Piscitelli, M.S., Capozzoli, A., 2025. Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of Heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings. *Appl. Energy* 381, 125046. <http://dx.doi.org/10.1016/j.apenergy.2024.125046>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. <http://dx.doi.org/10.48550/arXiv.1707.06347>, arXiv:1707.06347.
- Sendra-Arranz, R., Gutiérrez, A., 2020. A long short-term memory artificial neural network to predict daily HVAC consumption in buildings. *Energy Build.* 216, 109952. <http://dx.doi.org/10.1016/j.enbuild.2020.109952>.
- Shakeel, A., Chong, D., Wang, J., 2023. Load forecasting of district heating system based on improved FB-Prophet model. *Energy* 278, 127637. <http://dx.doi.org/10.1016/j.energy.2023.127637>.
- Somu, N., Raman M. R. G., Ramamritham, K., 2021. A deep learning framework for building energy consumption forecast. *Renew. Sustain. Energy Rev.* 137, 110591. <http://dx.doi.org/10.1016/j.rser.2020.110591>.
- Sutton, R.S., Barto, A.G., 1998. Reinforcement learning: An introduction. In: *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, Mass.
- Syed, D., Abu-Rub, H., Ghayeb, A., Refaat, S.S., 2021. Household-level energy forecasting in smart buildings using a novel hybrid deep learning model. *IEEE Access* 9, 33498–33511. <http://dx.doi.org/10.1109/ACCESS.2021.3061370>.
- Taylor, S.J., Letham, B., 2018. Forecasting at scale. *Amer. Statist.* 72 (1), 37–45. <http://dx.doi.org/10.1080/00031305.2017.1380080>.
- Wang, H., Wang, S., 2021. A hierarchical optimal control strategy for continuous demand response of building HVAC systems to provide frequency regulation service to smart power grids. *Energy* 230, 120741. <http://dx.doi.org/10.1016/j.energy.2021.120741>.
- Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., Dai, B., Miao, Q., 2024. Deep reinforcement learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* 35 (4), 5064–5078. <http://dx.doi.org/10.1109/TNNLS.2022.3207346>.
- Xia, Y., Wang, X., Yin, X., Bo, W., Wang, L., Li, S., Li, K., 2024. Federated accelerated deep reinforcement learning for multi-zone HVAC control in commercial buildings. *IEEE Trans. Smart Grid* <http://dx.doi.org/10.1109/TSG.2024.3524756>, 1–1.
- Xiao, Z., Gang, W., Yuan, J., Chen, Z., Li, J., Wang, X., Feng, X., 2022. Impacts of data preprocessing and selection on energy consumption prediction model of HVAC systems based on deep learning. *Energy Build.* 258, 111832. <http://dx.doi.org/10.1016/j.enbuild.2022.111832>.
- Yang, L., Yan, H., Lam, J.C., 2014. Thermal comfort and building energy consumption implications – A review. *Appl. Energy* 115, 164–173. <http://dx.doi.org/10.1016/j.apenergy.2013.10.062>.
- Yu, P., Zhang, H., Song, Y., Hui, H., Chen, G., 2024. District Cooling System Control for Providing Operating Reserve Based on Safe Deep Reinforcement Learning | IEEE Journals & Magazine | IEEE Xplore. <https://ieeexplore.ieee.org/document/10019581>.
- Zhang, C., Kuppannagari, S.R., Prasanna, V.K., 2022. Safe building HVAC control via batch reinforcement learning. *IEEE Trans. Sustain. Comput.* 7 (4), 923–934. <http://dx.doi.org/10.1109/TSUSC.2022.3164084>.
- Zhu, J., Tian, Z., Niu, J., Lu, Y., Cheng, B., Zhou, H., 2025. Machine learning-enhanced lightweight rule-based control strategy for building energy demand response. *Build. Simul.* 18 (7), 1857–1876. <http://dx.doi.org/10.1007/s12273-025-1275-1>.
- Zhuang, D., Gan, V.J.L., Duygu Tekler, Z., Chong, A., Tian, S., Shi, X., 2023. Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning. *Appl. Energy* 338, 120936. <http://dx.doi.org/10.1016/j.apenergy.2023.120936>.