Applied Thermal Engineering 295 (2026) 130741

Contents lists available at ScienceDirect

Applied Thermal Engineering

journal homepage: www.elsevier.com/locate/ate

Research Paper

Model predictive control guided imitation learning for optimal control of PCM thermal energy storage☆

Yangzhe Chen [a,*], Ilaria Marotta [b], Valeria Palomba [b], Thomas Ohlson Timoudas [c], Qian Wang [a,d]

[a] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm 10044, Sweden
[b] National Research Council (CNR) of Italy, Institute for Advanced Energy Technologies (CNR-ITAE), 98126 Messina, Italy
[c] RISE Research Institutes of Sweden, Division Digital Systems, Computer Science, Isafjordsgatan 22, Kista 164 40, Sweden
[d] Uponor AB, Hackstavägen 1, Västerås 721 32, Sweden

ARTICLE INFO

Keywords:
Imitation Learning
Model predictive control
PCM storage
Energy flexibility
Demand-side management

ABSTRACT

The integration of Phase Change Material (PCM) storage with Heat Pump (HP) systems offers significant potential for demand-side flexibility but presents challenges in control due to complex thermodynamics during phase change. To overcome the computational burden of online optimization and the training instability of model-free reinforcement learning, this study proposes a novel framework utilizing Model Predictive Control (MPC)-guided Imitation Learning (IL). A high-fidelity Functional Mock-Up Unit (FMU) is employed to simulate the PCM-HP integration, where an MPC expert agent generates optimal control trajectories. Two IL agents, Behavior Cloning (BC) and Generative Adversarial Imitation Learning (GAIL), are trained to mimic this expert under dynamic pricing signals. While both IL agents are able to learn the load-shifting behaviors, GAIL outperforms BC in generalization. BC suffers from limited robustness in unobserved states, whereas GAIL captures the underlying policy distribution, achieving Mean Absolute Percentage Error (MAPE) of approximately 9% during testing. This framework successfully bridges model-based and model-free paradigms, offering a scalable, real-time control alternative that retains optimality without requiring complex physical modeling during deployment.

1. Introduction

In recent decades, Europe has been experiencing increasingly frequent and intense heat waves, posing significant challenges to energy systems in terms of maintaining thermal comfort and ensuring energy stability [1]. Against this backdrop, advances in energy systems have emphasized the importance of leveraging Demand-Side Management (DSM) strategies to enhance energy flexibility. DSM has served as a crucial mechanism for aligning supply with demand, particularly in optimizing the operation of energy systems and integrating variable renewable energy sources [2]. While DSM has played a critical role in the heating sector, especially through low-temperature heating applications integrated with Thermal Energy Storage (TES) and HP, its potential for cooling remains underexplored [3–5]. This is particularly relevant for Mediterranean climates, where HPs serve as an alternative to conventional air conditioners. However, one of the key challenges is how to provide cooling economically, given Energy Efficiency Ratio (EER) and pricing signals, to unlock the energy flexibility potential. These trends have catalyzed the development and deployment of TES systems, which serve as buffers for thermal demand in buildings due to their low cost and long service life compared to other energy storage technologies [6]. Among various TES technologies, latent TES systems represent an advanced class that enhances the fundamental benefits of TES by utilizing both sensible and latent heat storage mechanisms, thereby increasing system efficiency and reducing spatial demand; an example is the application of PCM storages [7].

Recently, the control strategies of building-TES integration have evolved significantly, primarily from rule-based to model-based optimal control, e.g. MPC [8,9]. The optimization feature of MPC has gained significant attention in the building sector, due to its capability to predict future system states and optimize actions based on various boundary conditions [10,11]. However, the implementation of MPC is limited by its model-dependent nature. The performance of such control is determined by precise component models. Traditional physics modeling has been investigated extensively. However, this modeling

☆ This article is part of a Special issue entitled: 'ITT-BESS' published in Applied Thermal Engineering.

* Corresponding author.
E-mail address: yangzhec@kth.se (Y. Chen).

https://doi.org/10.1016/j.applthermaleng.2026.130741
Received 5 October 2025; Received in revised form 15 February 2026; Accepted 21 March 2026
Available online 23 March 2026

approach requires comprehensive domain knowledge and manufacturer parameters. In addition, it is also challenging to develop data-driven models for such complex physical systems [12]. This is mainly due to the unmeasurable state of the phase fraction and deviations from theoretical assumptions in real practice [13].

To address the aforementioned limitation, model-free control strategies have emerged as a promising tool to improve the energy efficiency and operational effectiveness of buildings, while reducing the efforts needed for model development and data collection. In recent years, model-free control approaches, such as Deep Reinforcement Learning (DRL), have demonstrated the potential in cooling applications by enabling adaptive and optimal control without relying on explicit system models [14]. These methods can manage complex thermodynamics and uncertainties in cooling systems, improving energy flexibility and occupant comfort [15]. A DRL agent can be trained either in a real environment or a simulated environment. However, both approaches tend to be time-consuming and unstable due to their trial-and-error feature. Moreover, the physical boundary conditions cannot be incorporated explicitly, and this may lead to extreme conditions, especially at the early stage of training in a real environment, for example, the condensation problem in cooling application [16]. Therefore, in order to facilitate the training process and adhere to physical boundary conditions, knowledge transfer methods are needed.

Imitation Learning (IL) is a method that allows an agent to learn to perform tasks by mimicking expert behavior rather than completely going through a trial-and-error process [17]. IL differs from traditional DRL, as it does not involve the agent learning solely from its own experiences. Instead, IL utilizes the actions demonstrated by an expert to guide the agent's behavior. In this process, the target agent accesses trajectories provided by the expert, which encompass a series of states or state–action pairs.

To address the challenges of achieving cost-effective cooling control in PCM and HP, this study proposes a hybrid solution that combines IL with MPC. The difference between MPC and MPC-guided IL is that while the former relies on iterative, online optimization requiring precise physical models and external solvers to handle system dynamics, MPC-guided IL shifts this computational burden to the offline training phase. By imitating the expert's control policy, the IL framework enables model-free, real-time execution that retains the expert's optimality without requiring complex physical parameters during deployment. In addition, compared to DRL, MPC-guided IL utilizes supervised learning to directly imitate the expert's trajectories instead of relying on time-consuming trial-and-error exploration to discover optimal policies. This allows the agent to bypass the unstable exploration phase and inherently adhere to physical boundary conditions demonstrated by the expert. This solution enables the agent to learn expert trajectories that are trained on simplified surrogate models, thereby reducing the dependence on detailed system models and extensive data collection. The novelty of this approach lies in bridging model-based and model-free control paradigms to deliver effective control strategies with reduced training complexity. Specifically, this method is applied to cool buildings using PCM and HP with minimum cost, while efficiently adhering to physical boundary conditions to ensure robust decision-making.

1.1. Previous studies

1.1.1. Building-TES integration

The optimal control of building-TES has been studied extensively from previous studies. Baniasadi et al. proposed a coordinated design and operation framework for TES in smart buildings, showing that their method can achieve more than an 80% reduction in annual electricity costs and over a 42% decrease in life cycle costs [18]. Bampoulas et al. introduced an ensemble learning-based framework to evaluate the energy flexibility of residential buildings equipped with TES, achieving high accuracy in day-ahead flexibility predictions with scores of 0.979 and 0.968 [19]. Chen et al. proposed a flexibility-centric framework that integrates data-driven approaches for optimal sizing and operation of TES in buildings. Their approach demonstrated up to 35% savings in operational costs and achieved capacity utilization efficiencies as high as 99% across varied building types [20]. The aforementioned studies mainly focus on sensible TES in buildings. On the one hand, Alghamdi et al. explored the use of a 5 cm PCM layer in building envelopes alongside a new multi-setpoint PID controller to enhance energy management, achieving a 15.3% reduction in annual electricity consumption and an additional 2.2% saving through optimal chiller control [21]. Finck et al. quantified the demand-side flexibility of building heating systems by integrating HP with PCM [22]. It showed that optimal control of TES could reduce operational electricity costs by up to 7.1%. Tan et al. conducted a techno-economic evaluation of a PCM-based TES system in an office building, showing that only 36% of the theoretical capacity was practically usable due to slow charging rates. Using optimal control, the system achieved a 4.2% reduction in annual cooling costs, with a maximum investment payback threshold of 921 € over five years [23].

1.1.2. Imitation learning in buildings

IL is a paradigm in DRL, which an agent learns to perform a task by supervised learning from expert demonstrations [17]. Silvestri et al. deployed an IL controller using BC method in a real office building Heating Ventilation and Air Conditioning (HVAC) system. Their method achieved a 40% reduction in energy consumption and up to 43% fewer temperature violations compared to traditional Rule Based Control (RBC) controllers [24]. Liu et al. introduced an imitation-interaction learning method that combines BC with DRL to optimize multi-zone ventilation systems, achieving up to 18.9% energy savings and reducing training steps by 66.4% compared to the standard baseline [25]. Dinh and Kim introduced an optimization based IL framework for residential HVAC control that trains a deep neural network to mimic mixed-integer linear programming without relying on forecast data. Their method achieved near-optimal control with only 0.006 kW Mean Absolute Error (MAE) in hourly power usage, while maintaining thermal comfort [26]. Dengiz and Kleinebrahm developed a forecast-independent control approach for HP by integrating IL with a heuristic Price-Storage-Control algorithm. Their method reduced electricity costs by up to 4.7% with the integration of a neural network and IL [27]. Liu et al. evaluated GAIL for a fan coil unit control in a commercial building with the expert demonstration of a MPC controller, achieving 95% of expert-level cumulative reward while reducing energy costs by 21% and temperature violations by 92% compared to rule-based control [28]. Park et al. presented an IL strategy for passive control of shading, ventilation, and insulation in buildings, achieving over 90% reduction in space conditioning loads across three of four climate types. Their approach matched expert control with over 95% action accuracy, and generalized robustly across 24 cities and diverse dwelling configurations [29]. Dey et al. proposed an imitation learning–augmented DRL approach for HVAC control, using an EnergyPlus-based building model to stabilize early training and accelerate convergence. Their method reduced training time equivalent to four years of summer data and achieved a 6.3% reduction in energy cost and a 7.2% improvement in performance score compared to rule-based control [30]. Amasyali et al. proposed a transfer learning approach to enhance the data efficiency of DRL controllers for air conditioners, enabling knowledge transfer across buildings. Their method reduced training time by up to 60% while maintaining thermal comfort and generalizing effectively without full retraining [31]. Dey et al. proposed an inverse reinforcement learning framework to initialize DRL-based HVAC controllers using rule-based data, enhancing early training stability and sample efficiency. Their method achieved a 38% cost reduction over direct DRL training and surpassed both offline and metamodel-based approaches in managing energy use and thermal comfort under demand response conditions [32]. The key findings from the literature are summarized in Table 1.

Table 1
Summary of previous studies on IL in buildings.

Reference	IL approach	Expert source	Application	Key finding
[24]	BC	Measured (RBC)	Hydraulic heating	40% energy reduction
[25]	BC	Measured (RBC)	Multi-zone ventilation	18.9% energy savings
[26]	BC	Simulated (MPC)	Air conditioner	0.1 °C in temperature deviation
[27]	BC	Simulated (MPC)	HP	4.7% cost reduction
[28]	GAIL	Simulated (MPC)	Fan coil unit	21% energy cost reduction
[29]	BC	Measured (RBC)	Passive cooling	>50% load reduction
[30]	BC	Simulated (RBC)	Cooling coil	6.3% energy cost reduction
[31]	Transfer learning	Measured (RBC)	Air conditioner	60% training time reduction
[32]	Inverse RL	Simulated (RBC)	Cooling coil	38% cost reduction

1.1.3. Research gap

Although IL has been shown promising in accelerating training and enhancing policy robustness for building HVAC control, limited research has examined its application in PCM-HP cooling systems. Moreover, limited research has examined the integration of IL with physically constrained, model-based expert agents, such as MPC in the context of cooling applications.

1.2. Contribution

Based on the research gaps mentioned in the previous section, this study presents a novel framework that bridges model-based and model-free control agents that reduces the complexity in system modeling. By generating expert control actions from a reduced-order model based MPC, the key contributions of this work are summarized as follows:

- A high-fidelity FMU is developed for a PCM storage integrated HP system for cooling. The surrogate model is based on a validated fast reduced-order model and enables co-simulation with control agents.
- MPC is formulated to optimize the operation of the integrated system by explicitly incorporating physical constraints and a trade-off between operational cost and the mechanical stress of the compressor. A Pareto frontier analysis is conducted to determine the optimal penalty weight ω for balancing these objectives.
- Two IL approaches, BC and GAIL, are trained using the augmented expert policy generated from MPC. In addition, the performance of them in the peak cooling season is evaluated and compared.
- The proposed method is tested using real data from an office building located in the Mediterranean climate.

2. Methodology

This section introduces the methodology adopted in this study to develop and evaluate a hybrid control framework combining MPC the expert agent with IL for PCM-HP cooling systems. An overview of the methodology is shown in Fig. 1, and the framework is divided into three phases.

Phase 1. This phase develops the simulation environment and formulates the expert agent. Specifically, the building cooling load is introduced in Section 2.1.1. To simulate the renovation of the building, that is, integrating the PCM storage with HP, an FMU is developed and validated against experimental data, as described in Section 2.1.2. In addition, an MPC controller is created as an expert agent using a reduced order surrogate model, the formulation of which is discussed in Section 2.2.

Phase 2. This phase runs the MPC controller in the simulation environment to generate expert control trajectories. The detailed processing of the expert control trajectories is introduced in Sections 2.3.1 and 2.3.2.

Phase 3. In the last phase, two different IL agents are trained using the processed control trajectories: BC and GAIL, as detailed in Sections 2.3.3 and 2.3.4. Moreover, the IL agents are implemented in the simulation environment for performance evaluation, and the evaluation metrics are elaborated in Section 2.4.

2.1. System modeling

This section outlines the simulation environment, consisting of building cooling load and PCM-HP FMU modeling.

2.1.1. Building cooling load

The use case is an office building located in the Mediterranean area of Italy. The building includes several office rooms, and other zones used as bathrooms and a corridor. The building cooling system is an HP with a thermal capacity of 20 kW. The building is modeled and simulated in non-steady state conditions, in the TRNSYS environment, with time steps of 15 min and for a simulation period of 1 year [33]. The simulation results are calibrated and validated from real measurements of the building's electricity consumption. For more details on the model, readers can refer to [34]. As a benchmark, the original cooling load, without PCM storage, is considered as the baseline scenario. The unit electricity consumption for cooling demand and outdoor temperature are plotted in Fig. 2(a). In addition, the unit electricity consumption for cooling demand and electricity prices for peak cooling season (i.e., July and August) are depicted in Fig. 2(b).

2.1.2. PCM-HP FMU

The renovation to the target building is proposed as coupling PCM storage with an HP system, shown in Fig. 3. Such a combination has been tested at the National Research Council (CNR) - Institute for Advanced Energy Technologies (ITAE). In particular, a prototype is built to be operated as a small-size reversible HP. The main feature of the HP is the use of a low-Global Warming Potential (GWP) refrigerant, i.e. R1234Ze(E), which is commonly employed in the Organic Rankin Cycle, thus allowing to consider the future use as a dual-mode system (i.e. reversible HP). The HP is tested under variable speeds of the compressor to evaluate the possibility of having a flexible operation through control of partial load. More details of the experiment can be found in [35]. The main results of the experiment are shown in Fig. 4. The cooling power varies linearly with the condenser inlet temperatures and the compressor speed. The EER, defined as the ratio of cooling power to the electric power consumption of the HP, ranging from 8 to 5 across the experiment. The point at 600 rpm indicates that the cooling power increases with an increase in the condenser inlet temperature. The reason for such different behavior is the unstable operation of the compressor at the selected compressor speed, which results in a discontinuous cooling power. Since each point is the average of at least a one-hour test, the overall cooling power average at 25 °C is lower than the other points.

A distinct behavior is observed regarding the compressor speed. For speeds exceeding 50% of the maximum speed (N_{max} = 2900 rpm), the EER becomes independent of the compressor speed. Conversely, at lower part-load ratios than 50%, the efficiency drops significantly.

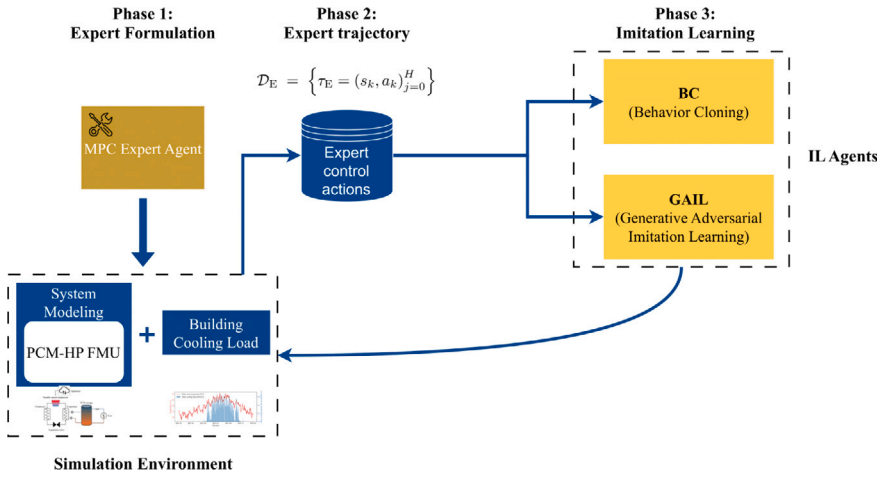
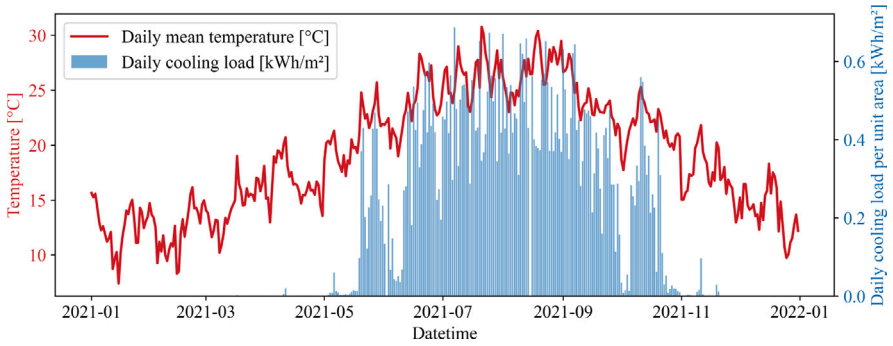
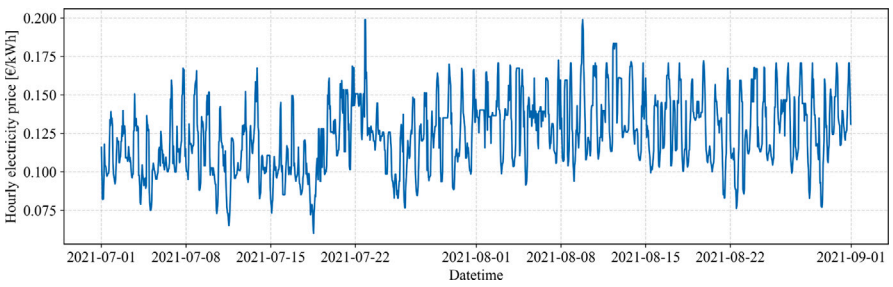


Fig. 1. Schematic overview of the IL framework. In Phase 1, the expert agent (MPC) and the simulation environment are developed. In Phase 2, expert control actions are collected by running MPC in the simulation environment. In Phase 3, the expert control actions are utilized for training the IL agents.



(a) The annual daily cooling load per unit area and the peak demand period during July and August, as indicated by the dashed box.



(b) The hourly electricity price during peak cooling season.

Fig. 2. The outdoor temperature, cooling load of the building, and electricity price from ENTSOE.

The HP model is developed in Modelica language within Dymola environment and is fully validated using experimental data [36]. Specifically, the numerical model utilizes the TIL library to simulate transient thermal behavior, defining the heat exchangers (condenser and evaporator) as parallel flow plate models with fixed UA values derived from experimental characterization. The compressor is represented by an efficiency-based model that relies on empirically determined volumetric and isentropic efficiencies, while the expansion valve operates as an orifice governed by a PI controller to regulate superheating [35]. The layout of the model of the HP in Dymola is shown in Fig. 5. It consists of the models for the plate heat exchangers for the (1) condenser and (2) evaporator, (3) the variable speed compressor, (4) the expansion valve. In the heat transfer fluid circuit (blue lines), there are valves that allow the use of the evaporator or the connection to (5) a storage (either a sensible storage or PCM storage).

The PCM storage model is also developed in Dymola and is based on the storage material and layouts discussed in [37]. The storage material used is a commercial salt hydrate with nominal melting temperatures of 17 °C (PlusICE S17) [38]. The storage is a metallic container, filled with FlatICE capsules of the selected storage materials.

The layout of the model for the PCM storage in Dymola software is presented in Fig. 6. It consists of the parallel connection of "PCM base" cells, which correspond to the model of a portion of the heat exchanger. This, in turn, includes the volume for the storage material, the walls that represent the external walls of the PCM capsules, and the liquid passages for the heat transfer fluid.

The model for the PCM-HP is exported as FMU, version 2.0, including the Dymola solver, to allow proper dynamic simulation of complex systems of partial differential equations [39]. The following parameters are used as inputs for the FMU: inlet temperatures of condenser and

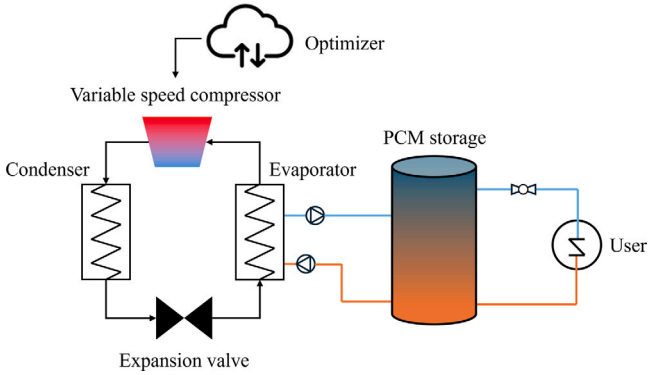


Fig. 3. The renovated PCM-HP system configuration.

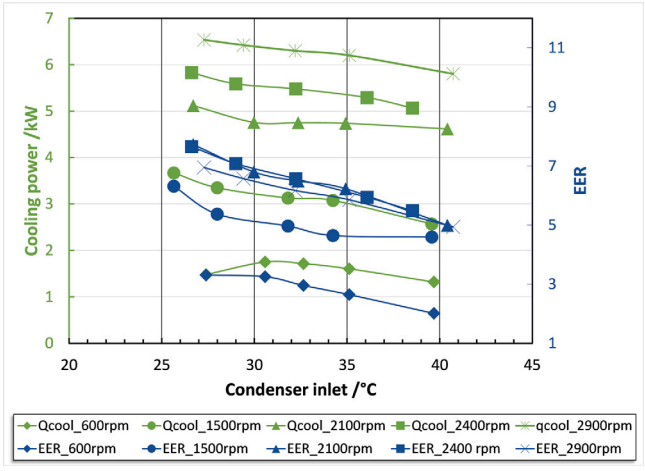


Fig. 4. The experimental results of the heat pump tested at CNR.

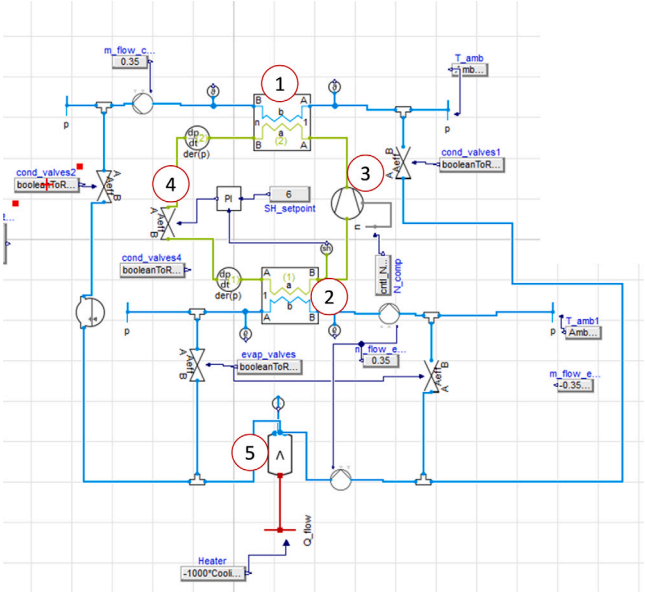


Fig. 5. The Dymola layout of the HP, (1) plate heat exchanger for the condenser, (2) evaporator, (3) variable speed compressor, (4) expansion valve, and (5) a storage.

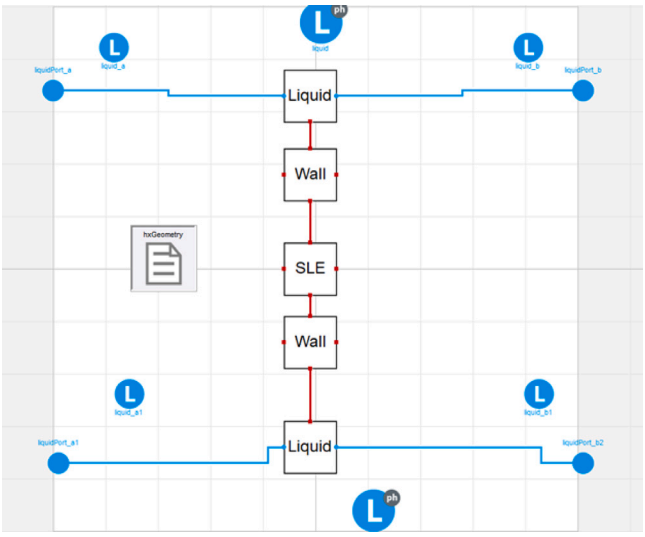


Fig. 6. The model for the PCM cell in the PCM storage model.

evaporator, inlet flow rates for condenser and evaporator, operating mode. The connection to the remaining part of the model is guaranteed by the output of the FMU, i.e. the temperatures of the heat transfer fluid in the condenser and evaporator and the inlet/outlet temperature of the PCM storage.

In addition, the difference in storage capacity between charging and discharging phases is necessary. This is because the simplified models that ignore hysteresis overestimate the actual storage capacity and result in poor accuracy in representing real world thermal behavior [40]. Therefore, in order to simulate the hysteresis of PCM, the capacity is considered 32 kWh during charging, and 27 kWh during discharging, based on the experiment measurement.

2.2. Expert formulation

In this paper, an MPC controller is developed to generate expert control actions.

2.2.1. Reduced-order modeling

The implementation of a reduced-order model is essential for MPC to mitigate the computational burden of high-fidelity physical simulations, therefore ensuring that the optimization problem remains tractable within the control horizon. It is worth noting that it is possible to use different reduced-order modeling approaches. In this study, quadratic polynomial regression is used to fit the electricity consumption of HP e^{hp}, cooling power Q^{hp} outdoor temperature T^{o}, and the compressor speed r, as shown in Eqs. (1) and (2). These models are fitted from a wide range of experimental data, by trying all the polynomial expressions up to two degrees and apply the best. The experiment data is split into training and test, with the ratio of 50/50. In addition, the R^2 and MAPE of the reduced-order models are presented in Table 2, demonstrating the strong correlation with the experimental data.

$$e_k^{\mathrm{hp}} = f(r_k, T_k^{\mathrm{o}}) = 1.5\times10^{-2} + 1.2\times10^{-4}\cdot r_k - 9.0\times10^{-3}\cdot T_k^{\mathrm{o}} - 1.5\times10^{-7}\cdot r_k^2 + 2.3\times10^{-5}\cdot r_k T_k^{\mathrm{o}} + 1.1\times10^{-3}\cdot (T_k^{\mathrm{o}})^2 \quad (1)$$

$$Q_k^{\mathrm{hp}} = g(r_k, T_k^{\mathrm{o}}) = -2.07\times10^{-1} + 5.33\times10^{-3}\cdot r_k - 1.37\times10^{-1}\cdot T_k^{\mathrm{o}} - 7.00\times10^{-7}\cdot r_k^2 + 1.24\times10^{-3}\cdot (T_k^{\mathrm{o}})^2 \quad (2)$$

Table 2
The accuracy of the reduced-order models.

Model	R^2	MAPE
Eq. (1)	0.99970	8%
Eq. (2)	0.99994	11%

2.2.2. MPC formulation

As the main feature of IL, the learned policy is directly determined by the behavior of the expert trajectory. In this study, the control sequence from an MPC agent is considered as the expert trajectory because it optimizes the control sequence based on the prediction of system dynamics and physical constraints. Employing the control sequence from MPC as the expert trajectory could effectively guide the learning process by providing high-quality demonstrations that inherently follow operational constraints and minimize costs. The formulation of MPC is shown in Eq. (3), with each of the constraints and symbols explained further down.

$$\min_{r_k} \quad \sum_{k=t}^{t+N-1} \left(\varepsilon_k \cdot e_k^{\mathrm{hp}} + \omega \cdot \|\Delta r_k\|_2^2 \right) \tag{3a}$$

$$\text{s.t.} \quad e_k^{\mathrm{hp}} := f(r_k, T_k^{\mathrm{o}}), \quad \forall k \in \{0, 1, \dots, N-1\} \tag{3b}$$

$$0 \le \mathrm{SoC}_k \le 1, \quad \forall k \in \{0, 1, \dots, N-1\} \tag{3c}$$

$$u_k^{\mathrm{dis}} = P_k^{\mathrm{load}}, \quad \forall k \in \{0, 1, \dots, N-1\} \tag{3d}$$

$$u_k^{\mathrm{ch}} \ge 0, \quad u_k^{\mathrm{dis}} \ge 0, \quad \forall k \in \{0, 1, \dots, N-1\} \tag{3e}$$

$$\Delta r_k = \{r_{k+1} - r_k, \dots, r_{k+N-1} - r_{k+N-2}\} \tag{3f}$$

$$0 \le r_k \le 2900, \quad \forall k \in \{0, 1, \dots, N-1\} \tag{3g}$$

Where:

- Eq. (3b) emphasizes the correlation between electricity consumption of HP (e_k^{hp}) and both the compressor speed (r_k), and the outdoor temperature (T_k^{o}), as described in Eq. (1).
- Eq. (3c) specifies the inequality constraint of State of Charge (SoC) of the PCM storage.
- Eq. (3d) represents that the supplied energy from PCM storage (u_k^{dis}) should be equal to the thermal load of the building (P_k^{load}).
- Eq. (3e) restricts the non-negativity of the charging (u_k^{ch}) and discharging energy (u_k^{dis}) of the storage.
- Eq. (3f) is the time series of compressor speed changes (Δr_k).
- Eq. (3g) defines the maximum compressor speed of the HP.

The objective function shown in Eq. (3a) is the sum of energy costs ($\varepsilon_k \cdot e_k^{\mathrm{hp}}$) and a l_2-norm based control action skew rate ($\|\Delta r_k\|_2^2$). For the first term, the electricity consumption of HP and electricity prices are denoted as e_k^{hp} and ε_k, respectively. In the latter term, r_k represents the control variable– compressor speed at time instance k, and ω is the weight to balance between two conflicting objectives. The control horizon is represented by N.

Since T_k^{o} is a known forecast value, $f(\cdot)$ becomes a quadratic function of the decision variable r_k, and can be directly substituted into the objective function (Eq. (3a)). Thus, Eq. (3b) is not a constraint, but a model-based definition aiming to formulate a convex quadratic objective. Likewise, it is evident that the optimization problem is a convex quadratic programming with a quadratic objective function and affine constraints. Additionally, the control horizon (N) is set as 12 h in this study and the time step is defined as 1 h. In order to solve the optimization problem, it is modeled in the Python package `cvxpy` and solved using the standard commercial solver MOSEK [41,42].

The MPC controller is simulated for August 2021 to collect training data for IL. A 12-hour prediction horizon is chosen to capture the daily load patterns between working and non-working periods, enabling the controller to schedule the intraday load-shifting. The time step of 1 h is selected to align with the temporal resolution of the day-ahead electricity market [43]. The two objectives are inherently conflicting, that is, minimizing operational cost often requires aggressive control actions changes. To balance this trade-off, the coefficient ω is introduced, weighting each objective in the objective function. A Pareto frontier analysis is performed to determine the optimal value of ω by repeatedly running simulations with ω values ranging from 0 to 20. The two objectives, namely operational cost and the average rate of change in compressor speed, are then compared. The optimal ω value is selected using the elbow method, which identifies the point on the Pareto curve where further improvement in one objective leads to only a marginal trade-off in the other [44].

Table 3
State space details of the expert trajectory.

States	Range	Description
$T^{\mathrm{o}}_{k,k+1,\dots,k+N-1}$	[10, 38]	Outdoor temperature from k to $k+N-1$
SoC_k	[0, 1]	SoC of the storage at time k
$\tilde{\varepsilon}_{k,k+1,\dots,k+N-1}$	[0, 1]	Normalized electricity price from k to $k+N-1$
$P^{\mathrm{load}}_{k,k+1,\dots,k+N-1}$	[0, 20]	Cooling load from k to $k+N-1$
$\tilde{r}_{k-1}$	[0, 1]	Normalized HP compressor speed at $k-1$

To account for prediction uncertainties, Gaussian noises ξ_k are added to the outdoor temperature (T^{o}) and building thermal load (P^{load}) at each step k of the prediction horizon. The noise follows a normal distribution $\xi_k \sim \mathcal{N}(0, \sigma_k^2)$, where the standard deviation σ_k increases linearly with the length of the prediction horizon N, as defined in Eq. (4):

$$\begin{aligned} \hat{x}_{t+k} &= x_{t+k} + \xi_k, \quad \text{where } \xi_k \sim \mathcal{N}(0, \sigma_k^2) \\ \sigma_k &= \sigma_{\mathrm{base}} + \alpha \cdot k, \quad \text{where } k \in [0, 1, 2, \dots, N-1] \end{aligned} \tag{4}$$

where, σ_0 represents the base forecast uncertainty, and α denotes the uncertainty growth rate per time instance k.

2.3. Imitation learning framework

The training of IL is based on the expert trajectory, which consists of a series of state–action pairs, as illustrated in Eq. (5) [45].

$$\mathcal{D}_{\mathrm{E}} = \left\{ \tau_{\mathrm{E}} = \left(s_k, a_k\right)_{j=0}^{H} \right\} \tag{5}$$

The state–action dataset is denoted as $\mathcal{D}_{\mathrm{E}}$, which comprises one or more expert trajectories τ_{E}. Each trajectory consists of a sequence of state–action pairs (s_k, a_k) collected from the start to the end of the simulation horizon H, with the time step of 1 h. The action space corresponds to the normalized compressor speed of the HP, bounded within the range [0, 1]. The state space is described in Table 3. It is important to note that the variables are normalized using the Min-Max scaling method.

2.3.1. Data augmentation

Data augmentation has been proven to be effective in enhancing the robustness of the model during IL [28,46]. Therefore, a zero mean Gaussian noise with a standard deviation of 0.05 is added to the normalized control actions in the expert trajectory. Fig. 7 shows the original control actions and the actions with augmentation.

2.3.2. Data separation and hyperparameter tuning

The augmented data is separated into three parts for training, validation, and testing. To be more specific, for model development, the model is trained and validated on state–action data collected in August, using an 80/20 random split, respectively. To ensure an unbiased evaluation, the fine-tuned model is then tested on data collected in July.

Hyperparameter tuning is conducted in the validation phase. In this study, the hyperparameter tuning leverages the Optuna framework [47]. This framework conducts automated hyperparameter optimization by iteratively proposing candidate parameter configurations, evaluating model performance for each trial, and utilizing the prior results to guide the search towards optimal values.

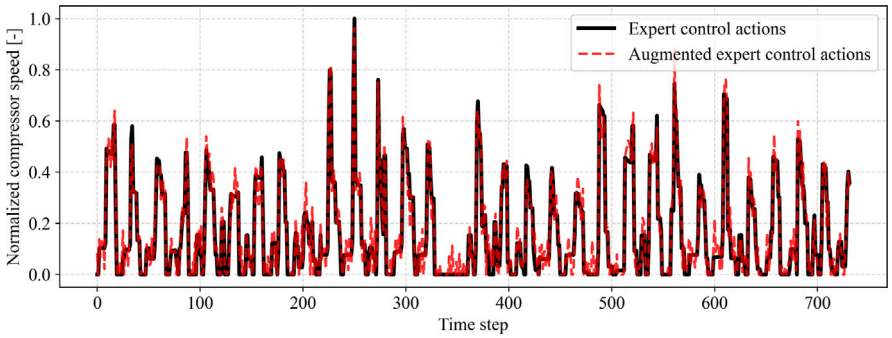


Fig. 7. The control actions in the expert trajectory and augmented by Gaussian noises with a standard deviation of 0.05 and a mean of 0.

Table 4
Results of hyperparameter tuning.

Hyperparameter	Candidate values	Step	Optimal
Batch size	[8, 128]	8	16
Weight decay	[1e−6, 1e−1]	–	5e−4
Learning rate	[1e−6, 1e−1]	–	4e−3
# of hidden layers	{1, 2, 3}	–	2
# of neurons	[16, 256]	8	Layer 1: 136 Layer 2: 232
Dropout rate	[0.0, 0.5]	0.02	0.08
Loss function	{'MSE', 'MAE', 'SmoothL1'}	–	'SmoothL1'
β	[0.1, 2.0]	0.1	1.94
Activate function	{'Relu', 'Tanh', 'Softmax'}	–	'Relu'

2.3.3. Behavior cloning

BC is a fundamental approach in IL. It is a supervised learning approach, guiding the agent to copy the expert demonstration [24]. In this study, a Multi-Layer Perceptron (MLP) is selected to train the BC agent. The candidate hyperparameters for validating the BC agent are presented in Table 4.

Three loss functions are evaluated as candidates: Mean Squared Error (MSE), MAE, and the SmoothL1loss function, as presented in Eqs. (6) and (7), respectively. The SmoothL1 loss introduces a hybrid criterion that applies a squared error term when the absolute element-wise prediction error is below a defined threshold β, and a l_1 term otherwise. This approach offers a balance between MSE and MAE, providing robustness to outliers and mitigating the risk of exploding gradients in some scenarios, as formulated in Eq. (8) [48]. Specifically, when the absolute difference between the predicted value x_n and the ground truth y_n exceeds a threshold β, the loss behaves identically to MAE; otherwise, a quadratic penalty is applied. To prevent overfitting, an early stopping strategy with maximum patience of 20 epochs is employed during the training process.

$$L_{\mathrm{MAE}}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N}\left|x_n - y_n\right| \tag{6}$$

$$L_{\mathrm{MSE}}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2 \tag{7}$$

$$\mathrm{SmoothL1}(\mathbf{x},\mathbf{y}) = \begin{cases} 0.5(x_n - y_n)^2/\beta, & \text{if } |x_n - y_n| < \beta \\ |x_n - y_n| - \frac{\beta}{2}, & \text{if } |x_n - y_n| \geq \beta \end{cases} \tag{8}$$

2.3.4. Generative adversarial imitation learning

The GAIL framework employed in this study is derived from the principles of Generative Adversarial Network (GAN) [49]. In contrast to BC, which attempts to replicate expert actions directly, GAIL aims to learn the underlying distribution of expert trajectories, thereby capturing more generalizable control behaviors [49].

The training process involves two neural networks, a generator and a discriminator, demonstrating the adversarial nature of the method. The generator is supposed to produce control actions that are similar to those in the expert trajectories. On the other hand, the discriminator works as a binary classifier, distinguishing between original expert actions and those generated by the generator. Hence, the generator is trained to 'fool' the discriminator by generating actions that are indistinguishable from expert trajectories, thereby enhancing the performance of policy through adversarial learning. The loss function of the discriminator is binary cross entropy, shown in Eq. (9), which is a common loss function in binary classification tasks.

$$L_D = -\mathbb{E}_{\tau_\pi}\left[\log D(s,a)\right] - \mathbb{E}_{\tau_\mathrm{E}}\left[\log(1 - D(s,a))\right] \tag{9}$$

Where, τ denotes the policy of the generator and τ_E represents the expert policy. The discriminator D is trained to assign high probabilities to expert state–action pairs and low probabilities to those generated by the learned policy. The first term penalizes the discriminator for incorrectly labeling generated actions as expert, while the second term penalizes it for failing to recognize true expert actions.

The loss function of generator in the GAIL framework is defined as:

$$L_G = -\mathbb{E}_{\tau_\pi}\left[\log D(s,a)\right] \tag{10}$$

This objective encourages the generator, which represents the learned policy τ_π, to produce actions in given states that the discriminator D cannot distinguish from those of the expert. It maximizes the likelihood that the discriminator classifies the generated state–action pairs (s,a) as expert-like.

Similar to BC, both generator and discriminator are modeled with MLP. In order to have a fair comparison, they both go through a hyperparameter optimization using Optuna as well.

2.4. Evaluation metrics

As mentioned in Sections 2.2 and 2.3, MPC and both IL are implemented in the simulation environment for comparison. To evaluate the performance of the control agents, the Flexibility Factor (FF) is used as an indicator [50]. It quantifies the capability of shifting load from high-penalty to low-penalty periods; in this case, it depends on the price signal. The expression of FF is shown in Eq. (11). In particular, the high-penalty and low-penalty periods are determined according to the daily average prices.

$$\mathrm{FF} = \frac{\int_{\text{low-penalty period}} P_\mathrm{h}\,dt - \int_{\text{high-penalty period}} P_\mathrm{h}\,dt}{\int_{\text{low-penalty period}} P_\mathrm{h}\,dt + \int_{\text{high-penalty period}} P_\mathrm{h}\,dt} \tag{11}$$

As sector coupling plays an important role in building energy flexibility, inspired from Pearson coefficient [51], a new indicator dedicated to thermal-electrical coupling is proposed in this study, as follows:

$$\mathrm{SCE} = \frac{\mathrm{cov}\left(\boldsymbol{P}^{\mathrm{nsh}},\, \boldsymbol{P}^{\mathrm{cool/heat}}\right)}{\sigma_{\boldsymbol{P}^{\mathrm{nsh}}} \cdot \sigma_{\boldsymbol{P}^{\mathrm{cool/heat}}}} \tag{12}$$

In Eq. (12), $\boldsymbol{P}^{\mathrm{nsh}}$ is the non-shiftable load time series (e.g., lighting and appliances), and $\boldsymbol{P}^{\mathrm{cool/heat}}$ is the cooling/heating load, which should be characterized by the same time resolution as $\boldsymbol{P}^{\mathrm{nsh}}$. In particular,

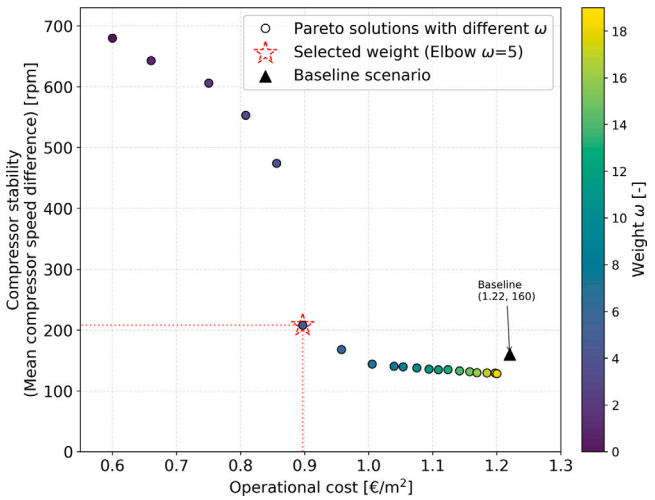


Fig. 8. Pareto frontier results for simulation in August. The black triangle represents the baseline scenario, and the red star denotes the selected value of ω due to the efficient balance between cost and compressor stability.

$\mathrm{cov}(\cdot)$ indicates the covariance operator, while σ denotes the standard deviation, in particular $\sigma_{\boldsymbol{P}^{\mathrm{nsh}}}$ and $\sigma_{\boldsymbol{P}^{\mathrm{cool/heat}}}$ for the respective series. The indicator measures the linear correlation between non-shiftable load (i.e., electricity consumption of appliances) and shiftable load (i.e., thermal load) and it ranges from −1 to +1. Values close to +1 indicate a positive correlation between the two loads, and it implies that the peak of both load occur at the same time. Conversely, values near −1 suggest a strong negative correlation, indicating that the thermal load is shifted to periods when the non-shiftable load is low, which demonstrates the performance of sector coupling.

3. Results

This section presents the tuning and simulation results of MPC controller and IL agents.

3.1. MPC tuning results

Fig. 8 presents the Pareto frontier, illustrating the trade-off between total operational cost and average HP compressor speed change rate. Each point represents a different simulation result associated with a specific penalty weight ω, indicated by the color scale on the right.

With weights increasing from 0 to 20, the average HP compressor change rates drop significantly. A distinct Pareto frontier is observed at the elbow point on the curve, corresponding to $\omega = 5$, indicating that increasing ω improves HP operation smoothness but also reduces operational costs. The average compressor speed change rate is reduced by around 400 rpm, compared to the optimization without the l_2-norm term. Additionally, in contrast to the baseline scenario, the operational cost is reduced by around 20%, despite a slight increase in the average compressor speed change rate of approximately 40 rpm.

3.2. MPC performance

After obtaining the most efficient weight from Pareto frontier analysis, Fig. 9 illustrates the overall performance of the MPC agent during the peak cooling season (i.e. July and August). As shown in Fig. 9(a), the weekly unit operational cost under the MPC control is consistently lower than that of the baseline scenario, achieving an approximate cost reduction of 20%. Fig. 9(b) compares the average cooling power between working hours (8 am–7 pm) and non-working hours. The results indicate that the MPC effectively shifts cooling demand towards non-working hours with lower electricity prices. Specifically, while the mean cooling power during the peak cooling season increases by 0.25 kW, it decreases by 0.27 kW during working hours. These findings demonstrate that the MPC strategy shifts cooling loads from high-price to low-price periods, thereby enhancing operational cost efficiency.

Table 5
Energy flexibility metrics comparison.

Metrics	Baseline	MPC	BC	GAIL
FF [–]	0.64	0.72	0.71	0.73
SCE [–]	0.75	0.65	0.62	0.64
Peak load [kW]	6.1	4.6	4.2	4.3
Seasonal EER [–]	6.8	7.3	7.3	7.3
Cost [€]	100.8	90.7	92.7	91.3

Note: [–] denotes dimensionless indicators; [kW] denotes kilowatts; [€] denotes Euros. FF: Flexibility Factor; SCE: Sector Coupling Efficiency; EER: Energy Efficiency Ratio.

3.3. Imitation learning training performance

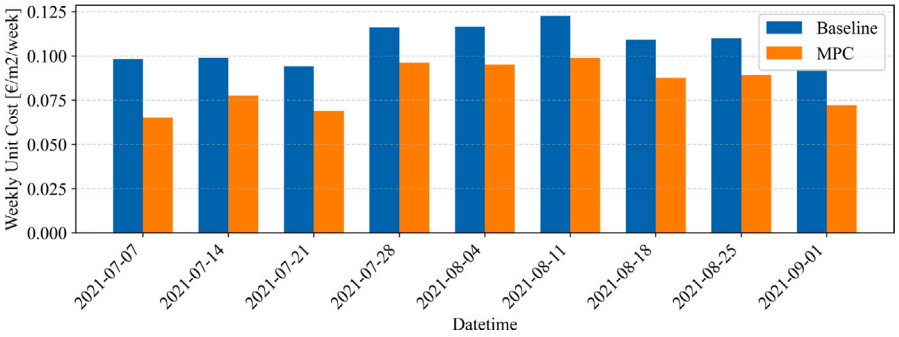
The control actions from MPC with ω of 5 is considered as the expert trajectory for both BC agent and GAIL agent training, as explained in Section 2.2.

BC training. Fig. 10 shows the training losses of BC agent. The chosen hyperparameters are presented in Table 4. The training process is terminated at approximately 70 epochs, as no significant improvement in the loss function is observed for 20 epochs, satisfying the early stopping criterion. A loss function (SmoothL1) of 0.002 is achieved for the training set, whereas it is 0.003 for the validation set and 0.064 in terms of Root Mean Squared Error (RMSE).

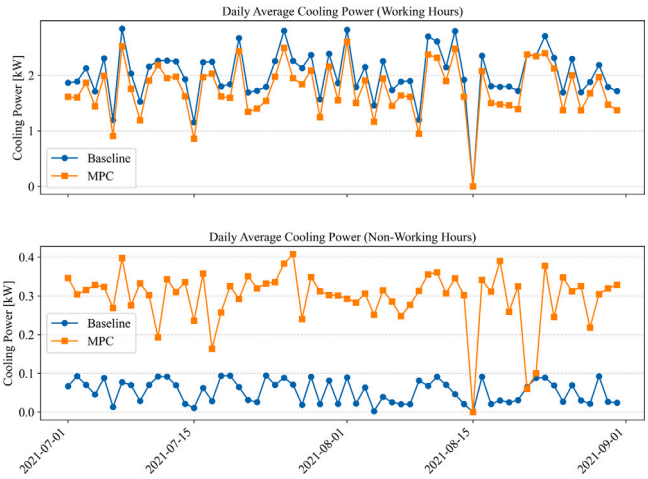
GAILTraining. The training performance of the GAIL agent is shown in Fig. 11. Due to the adversarial nature of the agent, the training fluctuates significantly in the first 300 epochs of training, however, both generator and discriminator converge after around 400 epochs. The generator loss settles around 0.65, while the discriminator loss stabilizes near 1.4. These values reflect a balanced state where the discriminator is no longer able to distinguish between expert and generated actions given the same states, and the generator produces actions that closely match the expert distribution. The discriminator loss being higher than 1.0 is typical in stabilized GAIL training, as it reflects a classification uncertainty due to overlapping distributions. Similarly, the generator loss remains below 1.0, suggesting that the policy has learned a behavior policy that closely mimics the expert demonstrations.

3.4. Comparison of energy and economic performance during test period

As introduced in Section 2.3, IL agents and the MPC controller are tested using the data from July. Fig. 12 compares the HP electricity consumption along with the pricing signals for all three agents. The top graph shows the performance during the training period, whereas the bottom one shows the test results of the first two weeks in the test period (i.e., 1–15 July). Overall, both IL agents demonstrate their capability to reproduce expert control policies. BC agent achieves MAPE of around 7% and 13% in August and July, respectively. In August, where cooling demand is higher due to elevated ambient temperatures, the BC agent achieves a near-identical consumption pattern to the MPC agent, reflecting its supervised learning ability to replicate expert control actions precisely. In contrast, the MAPEs of 9% and 10% are obtained for GAIL agent. Additionally, the GAIL agent displays a slightly smoother consumption profile compared to BC agent, especially around the peak cooling hours. Such behavior, while still aligning closely with the expert trajectory, results from its adversarial training objective to match the distribution of expert actions rather than replicating each action precisely.



(a) The weekly operational cost comparison between MPC and baseline scenario. The mean values of Baseline and MPC are 0.1 and 0.08 €/m^2/week, respectively. The standard deviation values are 0.0108 and 0.0127, respectively.



(b) The average cooling power comparison for both working hours (8 am–7 pm) and non-working hours. The average cooling loads of MPC and Baseline in working hours are 1.75 and 2.02 kW, and they are 0.3 and 0.05 kW for non-working hours, respectively.

Fig. 9. Overall performance of MPC agent during peak cooling season.

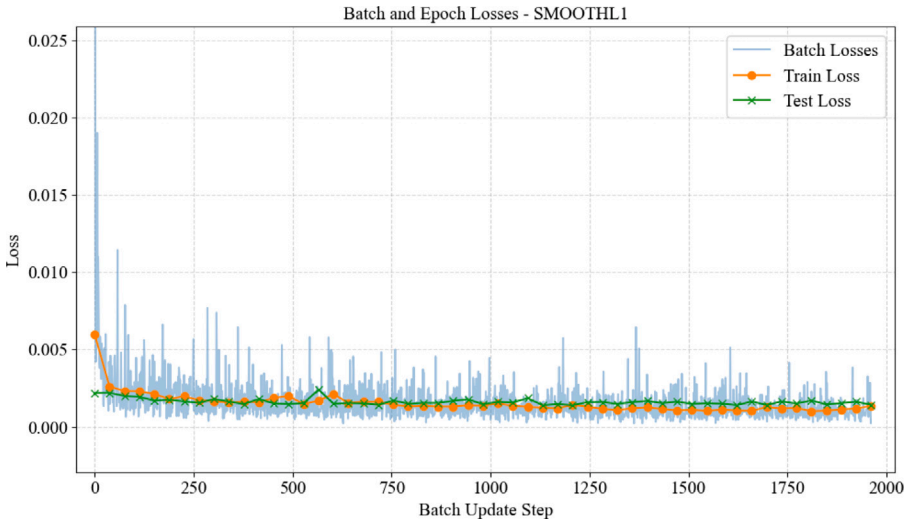


Fig. 10. Training losses of BC agent.

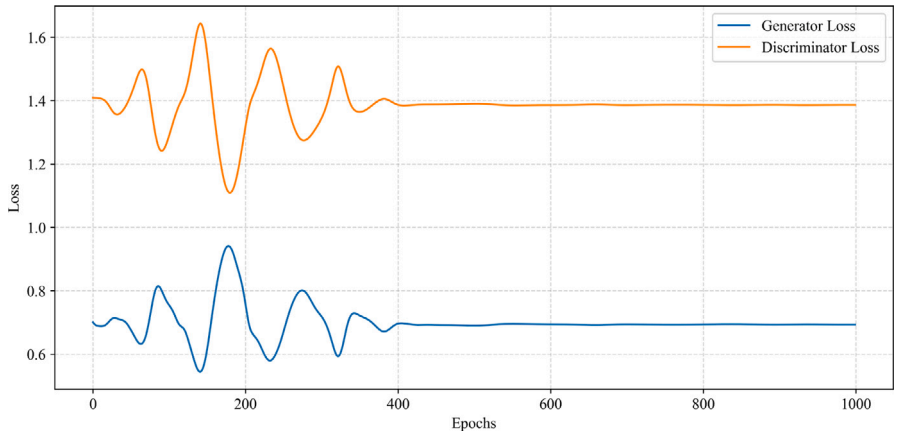
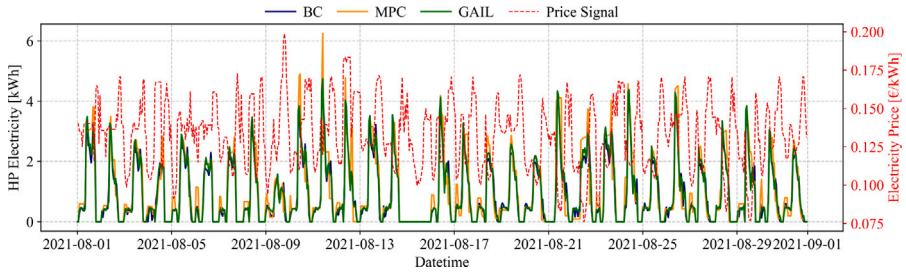
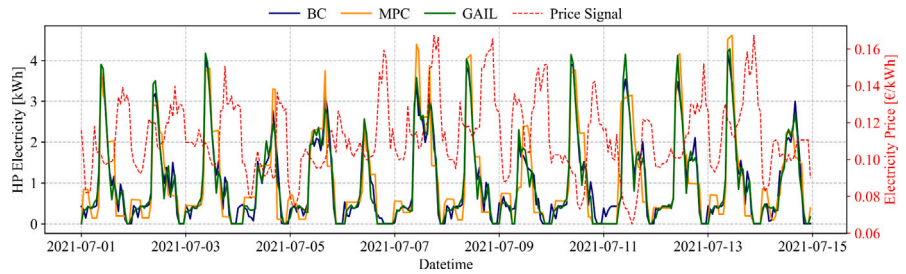


Fig. 11. Training losses of GAIL agent.



(a) The electricity consumption of HP and price in August.



(b) The electricity consumption of HP and price in July.

Fig. 12. Comparison of HP performance during July and August.

In addition to replicating the expert's control actions, Fig. 12 highlights the ability of the BC and GAIL agents to perform load-shifting in response to dynamic electricity pricing signals. Specifically, both agents demonstrate effective operation by aligning peak electricity consumption (e.g., charging PCM storage) with periods of relatively low pricing signals, thereby minimizing operational costs without compromising cooling supply. Table 5 presents the flexibility metrics comparison. It is worth noting that both IL agents achieve similar scores in these metrics, yet GAIL agent outperforms BC agent, demonstrating its superior imitation ability.

Fig. 13 depicts the cumulative operational cost during 1st and 15th July, in which the baseline scenario is represented by the black dashed line. Both BC and GAIL agents achieve a substantial reduction in operational cost by 11% and 13%, respectively. However, compared with MPC performance, the BC agent slightly deviates for around 2%, whereas GAIL shows almost the same reduction rate as expert.

Fig. 14 shows the average SoC of the PCM storage at each hour of the day during the validation period. The MPC agent charges the storage twice daily, aligning with the two valley pricing periods around 2–5 am and 12–3 pm, thereby optimizing operational costs through effective load-shifting. The GAIL agent mimics this pre-charging behavior in both periods. In contrast, although the BC agent also charges twice a day, it does not fully charge the storage during the afternoon low-price period (i.e. 12–3 pm), resulting in an increased usage of HP.

4. Discussion

This section contains the discussion about the performance comparison between two IL agents, as well as the limitation and future works.

4.1. Theoretical comparison between BC and GAIL

This study demonstrates that both BC and GAIL agents effectively mimic the expert MPC agent in operating the PCM-HP integrated system. However, the superior generalization capability of GAIL, particularly in maintaining lower policy errors than BC during testing, can be attributed to the fundamental theoretical differences in their learning objectives.

The BC agent treats the imitation problem as a supervised learning task, aiming to minimize the direct error between the expert's action

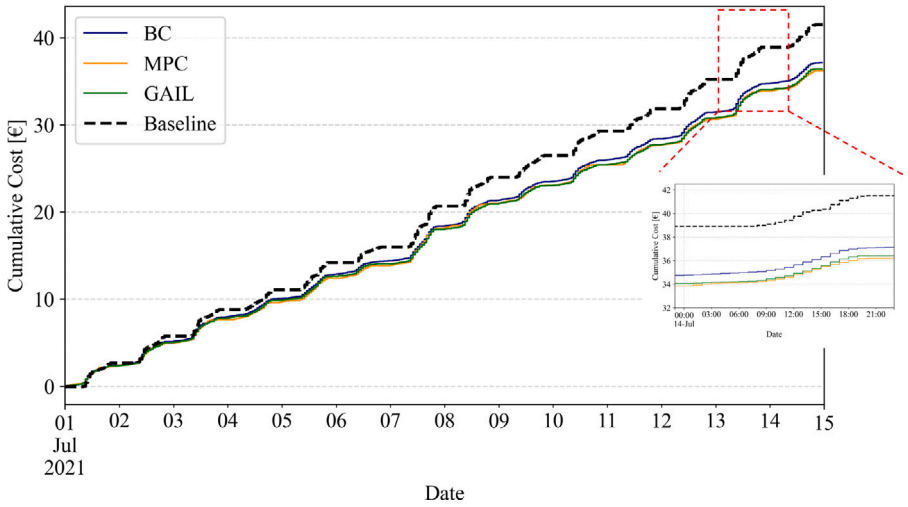


Fig. 13. The cumulative operational cost of HP in the validation period.

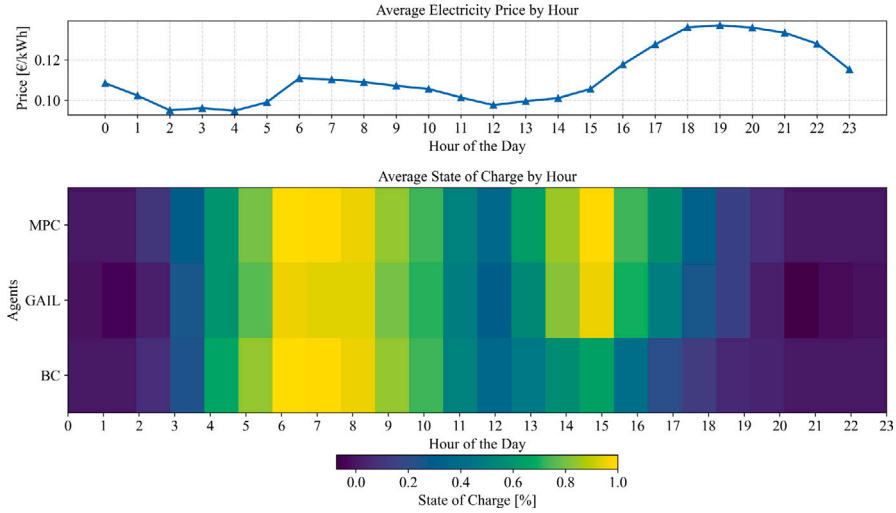


Fig. 14. The average SoC of the storage at each hour of a day.

a^* and the agent's action $a(s)$ given state s:

$$\mathcal{L}_{BC}(\theta) = \mathbb{E}\left[\ell(a(s), a^*)\right] \tag{13}$$

where ℓ is the loss function (e.g., SmoothL1). A critical limitation of this approach is the covariate shift problem. Since the agent is trained only on states visited by the expert, any slight deviation in the agent's action $a(s_t)$ leads to a next state s_{t+1} that may deviate from the expert's trajectory. As these deviations accumulate, the agent encounters states outside the training distribution, where the policy π_θ is limited.

In contrast, GAIL formulates imitation as a distribution matching problem, as expressed in Eqs. (9) and (10). Rather than minimizing point-wise action error, GAIL minimizes the Jensen–Shannon divergence between the occupancy measure of the expert τ_E and the learner τ_π.

By employing an adversarial discriminator D that distinguishes between expert and agent trajectories, the GAIL agent receives feedback not just on action accuracy, but on whether the resulting state–action distribution fits the expert's. This forces the agent to recover the expert's underlying intent rather than just memorizing specific actions. Theoretically, the error in GAIL scales linearly with the horizon, making it significantly more robust to compounding errors than BC. This explains why the GAIL agent consistently replicates the load-shifting strategy during unobserved validation periods (i.e., July), whereas the BC agent fails to fully charge the storage during the afternoon valley period, as shown in Fig. 14. That is due to the unseen states in the test data.

4.2. Computational performance

Table 6 shows the computational performance of both BC and GAIL. Although the total training time of both models is similar, around 2 minutes, the training time per epoch exhibits a significant discrepancy. While BC converges around 0.07 s per epoch, GAIL is relatively time-consuming, averaging 0.38 s per epoch. This is due to the adversarial nature of GAIL, that is, two neural networks need to be trained at the same time. Consequently, the computational performance plays a critical role when there is more data available in the future studies.

4.3. Performance in working and non/working hours

Fig. 15 plots the policy error distributions of the IL agents during working hours (8 am–6 pm) and non-working hours compared to the MPC agent. The corresponding mean values and standard deviations are summarized in Table 7. Overall, the GAIL agent demonstrates consistent performance across both periods, with negligible differences

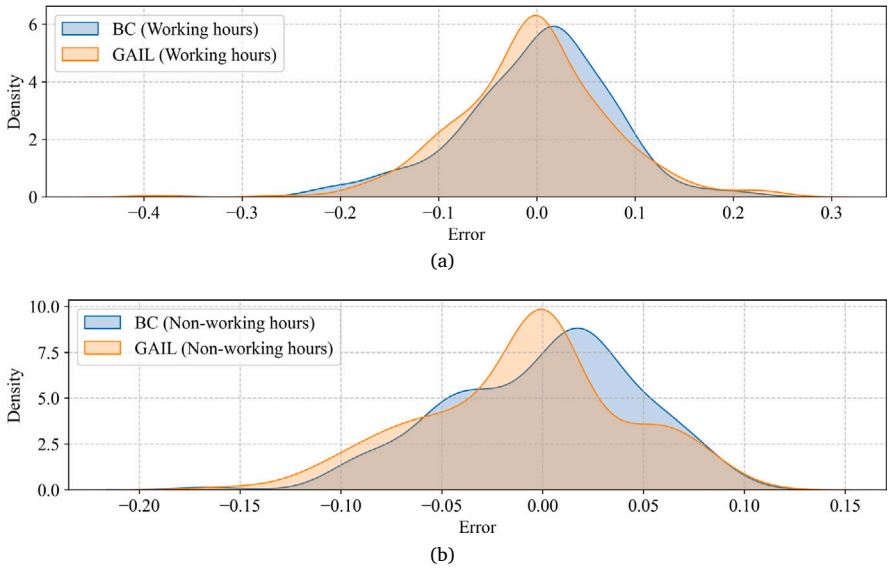


Fig. 15. The policy error distribution between IL agents and MPC agent during both working hours and non-working hours.

Table 6
The computational performance comparison between BC and GAIL. The total training time is determined using the selected combination from hyperparameter tuning.

	Training epochs [–]	Total training time [s]
BC	~1500	~110
GAIL	~400	~150

Table 7
The mean and standard deviations of policy error distributions.

	Mean	Standard deviation
BC (Working hours)	−0.01	0.08
BC (Non-working hours)	−0.01	0.05
GAIL (Working hours)	0.00	0.07
GAIL (Non-working hours)	0.00	0.04

in mean error and a 0.03 increase in standard deviation during working hours, indicating its robust generalization capability under varying operational conditions. In contrast, while the BC agent also maintains similar mean errors between periods, it has a relatively higher standard deviation compared to GAIL agent. This aligns with the average SoC of the storage at each hour of the day, as shown in Fig. 14. That is, during the first valley pricing signal period (i.e. 2 am–5 am, non-working hours), both IL agents manage to mimic expert policy to precharge the storage before pricing signal rises. However, during the second valley period (i.e. 12 pm–3 pm, working hours), while the GAIL agent still follows expert policy to charge the storage to nearly full, BC agent fails to imitate such behavior. Moreover, The higher standard deviation of policy errors observed during working hours for both agents can be attributed to the increased variability and complexity of building thermal loads and environmental conditions during these periods. Specifically, working hours typically coincide with peak load and dynamic pricing signals. These dynamic and rapidly changing conditions create a complex state space. As a result, both BC and GAIL agents exhibit greater variability in their control action errors when responding to these fluctuating conditions.

In this study, since the target building is an office building, the increment of flexibility metrics is limited due to no load during the night. As shown in Table 5, the score of FF metric is 0.64 in the baseline scenario, and the increment is only around 0.08 after the optimal control scenario is applied. Likewise, the reduction for Sector Coupling Efficiency (SCE) is around 0.1, attributing to the precharge behavior before working hours. On the other hand, there are substantial peak load reductions, especially during high pricing signal periods, by discharging the storage.

4.4. Limitation and future work

Despite the promising results demonstrated in this study, several limitations should be acknowledged. First, the current framework focuses solely on meeting supply–demand balance without explicitly considering indoor thermal comfort. In particular, the influence of relative humidity, which significantly affects perceived cooling performance and occupant comfort, is not incorporated in the simulation. Second, the datasets used for training and testing the MPC and IL agents correspond to July and August, which only represent the peak cooling season in the Mediterranean climate. While these months capture price and load fluctuations, they do not reflect the operational characteristics of other seasons with different temperature and cooling load distributions, i.e., mild cooling or transition seasons. As a result, the generalizability of the learned policies to cooling seasons in other years has not been assessed in this study.

Future work will address these limitations by integrating co-simulation with white-box building models to explicitly account for thermal comfort and humidity constraints. Moreover, the proposed IL framework can serve as an effective warm-up policy for further DRL training to keep exploring optimal and generalized control strategies beyond expert demonstrations, as pointed out in recent studies [24, 28]. The methodology can also be applied for heating-dominated PCM storages by replacing the reduced-order model of PCM. To summarize, the proposed methodology will be validated in other years when there is more data available, and a nested cross-validation is able to ensure the robustness of the performance.

5. Conclusion

This study proposes a novel framework that integrates Model Predictive Control (MPC) with Imitation Learning (IL) for optimal control of Phase Change Material (PCM)-Heat Pump (HP) integrated systems to provide demand-side energy flexibility. Two IL approaches, Behavior Cloning (BC) and Generative Adversarial Imitation Learning (GAIL), are trained using expert control actions generated by an MPC agent, and the performance in peak cooling season is compared. Although both IL

agents perform similarly in the training period, GAIL agent outperforms BC agent due to its adversarial training nature. The main findings can be summarized as follows:

- The MPC agent demonstrates operational performance by effectively operating the integrated system to minimize electricity costs in the context of integrating PCM-HP for cooling applications. The results show that a cost saving up to 20% is obtained via shifting the load to low pricing signal periods, and a peak load reduction of 1.5 kW is achieved compared to the baseline scenario.
- Both BC and GAIL agents are able to imitate the expert policy, achieving Mean Absolute Percentage Error (MAPE) of around 7% and 9% during the training phase, respectively. During training, BC agent slightly outperforms GAIL agent in replicating the expert's point-wise actions, indicating the importance of preparing comprehensive state–action pairs for BC training. In contrast, the GAIL agent exhibits superior generalization capability by learning the underlying distribution of expert trajectories, enabling robust policy performance under diverse and unobserved operational conditions.
- The proposed IL agents manage to learn the MPC policy while adhering to the physical boundary conditions and the load-shifting behavior. However, there are still discrepancies in performance, such as around 10% in policy errors and 2% in operational cost errors during the validation period. Future work will integrate co-simulation with white-box models to incorporate thermal comfort constraints and explore IL as an initialization for Deep Reinforcement Learning (DRL) to further optimize control strategies.

CRediT authorship contribution statement

Yangzhe Chen: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Ilaria Marotta:** Writing – review & editing, Software, Methodology, Data curation, Conceptualization. **Valeria Palomba:** Writing – review & editing, Validation, Software, Methodology, Data curation. **Thomas Ohlson Timoudas:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Qian Wang:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used [Gemini and ChatGPT] to improve the readability of the manuscript. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

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 Swedish Energy Agency, Sweden under Grant number 51544-1 and P2024-01185. This work was also European Union's Horizon Europe research and innovation program under Grant Agreement 101096789.

Appendix. Heat pump numerical modeling

The numerical model of the water-to-water reversible HP is developed within the Dymola environment using the Modelica language. The system components utilize the TIL library, which is specifically designed for the simulation of transient thermal systems. The model validation was performed against experimental data, achieving a deviation within 11% of the measured values. For more detail of the modeling, the readers can refer to [35].

A.1. Heat exchangers

The condenser and evaporator are modeled as parallel flow plate heat exchangers. The heat transfer process is characterized by a fixed global heat transfer coefficient (UA), which was determined from experiment. The heat transfer rate ($\dot{Q}$) is calculated based on the Logarithmic Mean Temperature Difference method:

$$\dot{Q} = UA \cdot \Delta T_{lm} \tag{14}$$

where ΔT_{lm} represents the logarithmic average of the temperature difference between the hot and cold inlets at each end of the heat exchanger:

$$\Delta T_{lm} = \frac{\Delta T_1 - \Delta T_2}{\ln(\Delta T_1 / \Delta T_2)} \tag{15}$$

Based on the experimental calibration, the UA values were set to 7132 W/K for the condenser and 11,379 W/K for the evaporator. A zero-pressure-drop assumption was applied for both the refrigerant and liquid sides.

A.2. Compressor

The compressor is simulated using an efficiency-based model governed by three constant efficiency parameters derived from experimental fitting: volumetric efficiency (η_{vol}), isentropic efficiency (η_{is}), and effective isentropic efficiency ($\eta_{eff,is}$).

The volumetric efficiency relates the actual mass flow rate ($\dot{m}_{act}$) to the theoretical mass flow rate ($\dot{m}_{theo}$):

$$\eta_{vol} = \frac{\dot{m}_{act}}{\dot{m}_{theo}} \tag{16}$$

The isentropic efficiency is defined as the ratio of the enthalpy increase during an ideal isentropic compression ($h_{is,out} - h_{in}$) to the actual enthalpy increase ($h_{out} - h_{in}$):

$$\eta_{is} = \frac{h_{is,out} - h_{in}}{h_{out} - h_{in}} \tag{17}$$

For this model, the validation shows constant values of $\eta_{vol} = 1$, $\eta_{is} = 0.55$, and $\eta_{eff,is} = 0.88$.

A.3. Expansion valve and control

The expansion device is modeled as an orifice valve with a variable effective area. The valve opening is regulated by a Proportional-Integral (PI) controller, which adjusts the flow area to maintain a fixed superheating set-point (SH_{set}) of 6 K at the evaporator outlet. The control logic minimizes the error $e(t)$ between the measured superheat and the setpoint:

$$e(t) = SH_{measured}(t) - SH_{set} \tag{18}$$

Data availability

The authors do not have permission to share data.

References

[1] European Environment Agency, European Climate Risk Assessment, Tech. Rep. EEA Report 01/2024, European Environment Agency, Copenhagen, Denmark, 2024, http://dx.doi.org/10.2800/867147, URL https://www.eea.europa.eu/publications/european-climate-risk-assessment. (Accessed 04 July 2025).

[2] I.A. 67, Energy in Buildings and Communities Programme Annex 67 Energy Flexible Buildings Summary Report, Tech. rep., International Energy Agency, 2019.

[3] T.J. Lindroos, J. Ikäheimo, Profitability of demand side management systems under growing shares of wind and solar in power systems, Energy Sources, Part B: Econ. Plan. Policy 19 (1) (2024) 2331487, http://dx.doi.org/10.1080/15567249.2024.2331487, arXiv:10.1080/15567249.2024.2331487.

[4] P.-H. Li, S. Pye, Assessing the benefits of demand-side flexibility in residential and transport sectors from an integrated energy systems perspective, Appl. Energy 228 (2018) 965–979, http://dx.doi.org/10.1016/j.apenergy.2018.06.153, URL https://www.sciencedirect.com/science/article/pii/S0306261918310237.

[5] M. Pantoš, L. Lukas, Enhancing power system reliability through demand flexibility of grid-interactive efficient buildings: A thermal model-based optimization approach, Appl. Energy 381 (2025) 125045, http://dx.doi.org/10.1016/j.apenergy.2024.125045, URL https://www.sciencedirect.com/science/article/pii/S0306261924024292.

[6] J. Liu, X. Yang, Z. Liu, J. Zou, Y. Wu, L. Zhang, Y. Zhang, H. Xiao, Investigation and evaluation of building energy flexibility with energy storage system in hot summer and cold winter zones, J. Energy Storage 46 (2022) 103877, http://dx.doi.org/10.1016/j.est.2021.103877, URL https://www.sciencedirect.com/science/article/pii/S2352152X21015449.

[7] A. Frazzica, M. Manzan, V. Palomba, V. Brancato, A. Freni, A. Pezzi, B.M. Vaglieco, Experimental validation and numerical simulation of a hybrid sensible-latent thermal energy storage for hot water provision on ships, Energies 15 (7) (2022) http://dx.doi.org/10.3390/en15072596, URL https://www.mdpi.com/1996-1073/15/7/2596.

[8] H. Tang, S. Wang, Energy flexibility quantification of grid-responsive buildings: Energy flexibility index and assessment of their effectiveness for applications, Energy 221 (2021) 119756, http://dx.doi.org/10.1016/j.energy.2021.119756, URL https://www.sciencedirect.com/science/article/pii/S0360544221000050.

[9] D. Wang, W. Zheng, Z. Wang, Y. Wang, X. Pang, W. Wang, Comparison of reinforcement learning and model predictive control for building energy system optimization, Appl. Therm. Eng. 228 (2023) 120430, http://dx.doi.org/10.1016/j.applthermaleng.2023.120430, URL https://www.sciencedirect.com/science/article/pii/S1359431123004593.

[10] D. Wang, Y. Chen, W. Wang, C. Gao, Z. Wang, Field test of Model Predictive Control in residential buildings for utility cost savings, Energy & Build. 288 (2023) 113026, http://dx.doi.org/10.1016/j.enbuild.2023.113026, URL.

[11] E. Saloux, J.A. Candanedo, C. Vallianos, N. Morovat, K. Zhang, From theory to practice: A critical review of model predictive control field implementations in the built environment, Appl. Energy 393 (2025) 126091, http://dx.doi.org/10.1016/j.apenergy.2025.126091, URL https://www.sciencedirect.com/science/article/pii/S0306261925008219.

[12] Q. Al-Yasiri, M. Szabó, Incorporation of phase change materials into building envelope for thermal comfort and energy saving: A comprehensive analysis, J. Build. Eng. 36 (2021) 102122, http://dx.doi.org/10.1016/j.jobe.2020.102122, URL https://www.sciencedirect.com/science/article/pii/S2352710220337542.

[13] T. Barz, D. Seliger, K. Marx, A. Sommer, S.F. Walter, H.G. Bock, S. Körkel, State and state of charge estimation for a latent heat storage, Control Eng. Pract. 72 (2018) 151–166, http://dx.doi.org/10.1016/j.conengprac.2017.11.006, URL https://www.sciencedirect.com/science/article/pii/S0967066117302642.

[14] B. Fina, H. Auer, W. Friedl, Profitability of active retrofitting of multi-apartment buildings: Building-attached/integrated photovoltaics with special consideration of different heating systems, Energy Build. 190 (2019) 86–102, http://dx.doi.org/10.1016/j.enbuild.2019.02.034, URL https://www.sciencedirect.com/science/article/pii/S0378778818333826.

[15] T. Wei, Y. Wang, Q. Zhu, Deep reinforcement learning for building HVAC control, in: Proceedings of the 54th Annual Design Automation Conference 2017, DAC '17, Association for Computing Machinery, New York, NY, USA, 2017, http://dx.doi.org/10.1145/3061639.3062224.

[16] D. Weinberg, Q. Wang, T.O. Timoudas, C. Fischione, A review of reinforcement learning for controlling building energy systems from a computer science perspective, Sustain. Cities Soc. 89 (2023) 104351, http://dx.doi.org/10.1016/j.scs.2022.104351, URL https://www.sciencedirect.com/science/article/pii/S2210670722006552.

[17] A. Hussein, M.M. Gaber, E. Elyan, C. Jayne, Imitation learning: A survey of learning methods, ACM Comput. Surv. 50 (2) (2017) http://dx.doi.org/10.1145/3054912.

[18] A. Baniasadi, D. Habibi, W. Al-Saedi, M.A. Masoum, C.K. Das, N. Mousavi, Optimal sizing design and operation of electrical and thermal energy storage systems in smart buildings, J. Energy Storage 28 (2020) 101186, http://dx.doi.org/10.1016/j.est.2019.101186, URL https://www.sciencedirect.com/science/article/pii/S2352152X19311545.

[19] A. Bampoulas, F. Pallonetto, E. Mangina, D.P. Finn, An ensemble learning-based framework for assessing the energy flexibility of residential buildings with multicomponent energy systems, Appl. Energy 315 (2022) 118947, http://dx.doi.org/10.1016/j.apenergy.2022.118947, URL https://www.sciencedirect.com/science/article/pii/S0306261922003646.

[20] Y. Chen, T. Ohlson Timoudas, Q. Wang, Flexibility-centric sizing and optimal operation of building-thermal energy storage systems: A systematic modelling, optimization and validation approach, Energy Build. 338 (2025) 115722, http://dx.doi.org/10.1016/j.enbuild.2025.115722, URL https://www.sciencedirect.com/science/article/pii/S0378778825004529.

[21] S.M. Alghamdi, M.N. Ajour, N.H. Abu-Hamdeh, A. Karimipour, Using PCM for building energy management to postpone the electricity demand peak load and approving a new PID controller to activate alternative chiller, J. Build. Eng. 57 (2022) 104884, http://dx.doi.org/10.1016/j.jobe.2022.104884, URL https://www.sciencedirect.com/science/article/pii/S235271022200897X.

[22] C. Finck, R. Li, R. Kramer, W. Zeiler, Quantifying demand flexibility of power-to-heat and thermal energy storage in the control of building heating systems, Appl. Energy 209 (2018) 409–425, http://dx.doi.org/10.1016/j.apenergy.2017.11.036, URL https://www.sciencedirect.com/science/article/pii/S0306261917316112.

[23] P. Tan, P. Lindberg, K. Eichler, P. Löveryd, P. Johansson, A.S. Kalagasidis, Thermal energy storage using phase change materials: Techno-economic evaluation of a cold storage installation in an office building, Appl. Energy 276 (2020) 115433, http://dx.doi.org/10.1016/j.apenergy.2020.115433, URL https://www.sciencedirect.com/science/article/pii/S0306261920309454.

[24] A. Silvestri, D. Coraci, S. Brandi, A. Capozzoli, A. Schlueter, Practical deployment of reinforcement learning for building controls using an imitation learning approach, Energy Build. 335 (2025) 115511, http://dx.doi.org/10.1016/j.enbuild.2025.115511, URL https://www.sciencedirect.com/science/article/pii/S0378778825002415.

[25] Y. Liu, Y. Song, C. Cui, Towards smart control and energy efficiency for multi-zone ventilation systems via an imitation-interaction learning method in energy-aware buildings, Energy 314 (2025) 134220, http://dx.doi.org/10.1016/j.energy.2024.134220, URL https://www.sciencedirect.com/science/article/pii/S0360544224039987.

[26] H.T. Dinh, D. Kim, MILP-based imitation learning for HVAC control, IEEE Internet Things J. 9 (8) (2022) 6107–6120, http://dx.doi.org/10.1109/JIOT.2021.3111454.

[27] T. Dengiz, M. Kleinebrahm, Imitation learning with artificial neural networks for demand response with a heuristic control approach for heat pumps, Energy AI 18 (2024) 100441, http://dx.doi.org/10.1016/j.egyai.2024.100441, URL https://www.sciencedirect.com/science/article/pii/S2666546824001071.

[28] M. Liu, M. Guo, Y. Fu, Z. O'Neill, Y. Gao, Expert-guided imitation learning for energy management: Evaluating gail's performance in building control applications, Appl. Energy 372 (2024) 123753, http://dx.doi.org/10.1016/j.apenergy.2024.123753, URL https://www.sciencedirect.com/science/article/pii/S030626192401136X.

[29] B. Park, A.R. Rempel, S. Mishra, Performance, robustness, and portability of imitation-assisted reinforcement learning policies for shading and natural ventilation control, Appl. Energy 347 (2023) 121364, http://dx.doi.org/10.1016/j.apenergy.2023.121364, URL https://www.sciencedirect.com/science/article/pii/S0306261923007286.

[30] S. Dey, T. Marzullo, X. Zhang, G. Henze, Reinforcement learning building control approach harnessing imitation learning, Energy AI 14 (2023) 100255, http://dx.doi.org/10.1016/j.egyai.2023.100255, URL https://www.sciencedirect.com/science/article/pii/S2666546823000277.

[31] K. Amasyali, Y. Liu, H. Zandi, A transfer learning strategy for improving the data efficiency of deep reinforcement learning control in smart buildings, in: 2024 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference, ISGT, 2024, pp. 1–5, http://dx.doi.org/10.1109/ISGT59692.2024.10454120.

[32] S. Dey, T. Marzullo, G. Henze, Inverse reinforcement learning control for building energy management, Energy Build. 286 (2023) 112941, http://dx.doi.org/10.1016/j.enbuild.2023.112941, URL https://www.sciencedirect.com/science/article/pii/S0378778823001718.

[33] Solar Energy Laboratory, University of Wisconsin-Madison, TRNSYS 18: A Transient System Simulation Program, Solar Energy Laboratory, University of Wisconsin-Madison, Madison, WI, USA, 2021, Version 18, [Online]. Available: https://sel.me.wisc.edu/trnsys/.

[34] I. Marotta, T. Péan, F. Guarino, S. Longo, M. Cellura, J. Salom, Towards positive energy districts: Energy renovation of a mediterranean district and activation of energy flexibility, Solar 3 (2) (2023) 253–282, http://dx.doi.org/10.3390/solar3020016, URL https://www.mdpi.com/2673-9941/3/2/16.

[35] O.A. Rehman, V. Palomba, A. Frazzica, A. Charalampidis, S. Karellas, L.F. Cabeza, Numerical and experimental analysis of a low-GWP heat pump coupled to electrical and thermal energy storage to increase the share of renewables across europe, Sustainability 15 (6) (2023) http://dx.doi.org/10.3390/su15064973, URL https://www.mdpi.com/2071-1050/15/6/4973.

[36] Dassault Systèmes, Dymola – Dynamic Modeling Laboratory, Dassault Systèmes AB, 2024, URL https://www.3ds.com/products-services/catia/products/dymola/. (Accessed May 2025), Version 2024.

[37] O.A. Rehman, V. Palomba, D. Verez, E. Borri, A. Frazzica, V. Brancato, T. Botargues, Z. Ure, L.F. Cabeza, Experimental evaluation of different macro-encapsulation designs for PCM storages for cooling applications, J. Energy Storage 74 (2023) 109359, http://dx.doi.org/10.1016/j.est.2023.109359, URL https://www.sciencedirect.com/science/article/pii/S2352152X23027573.

[38] PCM Products Ltd, PCM products - phase change material products and solutions, 2025, URL https://www.pcmproducts.net/. (Accessed 05 June 2025).

[39] Modelica Association Project FMI, Functional mock-up interface (FMI), 2025, URL https://fmi-standard.org/. (Accessed 05 June 2025).

[40] Y. Hu, R. Guo, P.K. Heiselberg, H. Johra, Modeling PCM phase change temperature and hysteresis in ventilation cooling and heating applications, Energies 13 (23) (2020) http://dx.doi.org/10.3390/en13236455, URL https://www.mdpi.com/1996-1073/13/23/6455.

[41] S. Diamond, S. Boyd, CVXPY: A python-embedded modeling language for convex optimization, J. Mach. Learn. Res. 17 (83) (2016) 1–5.

[42] MOSEK ApS, MOSEK Optimizer API for Python, MOSEK ApS, 2024, https://www.mosek.com.

[43] ENTSO-E, ENTSO-E transparency platform: Electricity generation and load data, 2025, URL https://transparency.entsoe.eu/. (Accessed 15 May 2025).

[44] E. Zitzler, L. Thiele, An evolutionary algorithm for multiobjective optimization: The strength Pareto approach, Tech. Rep.TIK-Report No. 43, ETH Zurich, 1998, URL https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/145900/1/eth-24834-01.pdf.

[45] B. Zheng, S. Verma, J. Zhou, I.W. Tsang, F. Chen, Imitation learning: Progress, taxonomies and challenges, IEEE Trans. Neural Networks Learn. Syst. 35 (5) (2024) 6322–6337, http://dx.doi.org/10.1109/TNNLS.2022.3213246.

[46] D. Antotsiou, C. Ciliberto, T.-K. Kim, Adversarial imitation learning with trajectorial augmentation and correction, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, 2021, pp. 4724–4730, http://dx.doi.org/10.1109/ICRA48506.2021.9561915.

[47] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.

[48] R. Girshick, Fast R-CNN, in: International Conference on Computer Vision, ICCV, 2015.

[49] J. Ho, S. Ermon, Generative adversarial imitation learning, 2016, arXiv:1606.03476, URL https://arxiv.org/abs/1606.03476.

[50] J. Le Dréau, P. Heiselberg, Energy flexibility of residential buildings using short term heat storage in the thermal mass, Energy 111 (2016) 991–1002, http://dx.doi.org/10.1016/j.energy.2016.05.076, URL https://www.sciencedirect.com/science/article/pii/S0360544216306934.

[51] K. Pearson, Notes on regression and inheritance in the case of two parents, Proc. R. Soc. Lond. 58 (1895) 240–242, http://dx.doi.org/10.1098/rspl.1895.0041.