International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 12, No. 4, 2025, pp. 469-486.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Thermodynamic-Aware Multi-Agent Reinforcement Learning with Predictive Load Shifting for Real-Time Power Optimization and Climate Control in Phase-Change Cold Storage Systems

Manasa G.V. Kumar [a], Dr.P.S. Rajakumar [b]

[a] *Research Scholar, Department of Computer Science and Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India. E-mail: manasagvkumar3@gmail.com, Orcid: https://orcid.org/0009-0005-3349-2894*

[b] *Professor, Department of Computer Science and Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India.*

Abstract

Phase-change cold storage systems have received serious consideration as a solution to peak reduction of electricity demand through the storage and release of thermal energy. Nevertheless, the intricate thermodynamic characteristics, particularly in the phase transitions, have inhibited their incorporation with reinforcement learning (RL) methods. The presented paper proposes a Thermodynamic-Aware Multi-Agent Reinforcement Learning (TA-MARL) framework, where physics-aware thermodynamic constraints, including latent heat and phase fraction dynamics, are directly embedded in the RL model. The methodology involved a hierarchical multi- time scale agent architecture, which allows predictive load shifting and energy arbitrage with regard to dynamic electricity prices and cooling demand predictions. The proposed system is compared to the baseline control strategies on the basis of real-world building energy and PCM thermodynamic simulations. The findings indicate that TA-MARL is superior to the conventional rule-based control (RBC) and price-only RL in attaining 27-32% savings in electricity costs and 22-28% savings in peak demand. Also, TA-MARL does not exceed safe limits of the phase transitions in the PCM, like in other strategies, and PCM constraints are not violated throughout the strategy. TA-MARL hierarchical organization is 35%-fold faster in converging on the learning than single-agent RL models. The ability to shift the predictive load greatly diminishes the use of energy under peak hours, and it also makes sure that the variation of temperature indoors is not below -0.5 o C and above +0.5 o C. To sum up, thermodynamic physics applied to RL allows creating a safer, efficient, and economically feasible system of PCM-based cold storage. The paper has also identified the significance of thermodynamic awareness in the reinforcement learning of thermal energy storage, and prepares the future study of extending this model to a larger, multi-zone system.

Keywords: Reinforcement Learning, Thermodynamic-Aware Multi-Agent, Load Shifting, Cold Storage Power Optimization, Phase-Change Material.

DOI: 10.71086/IAJSE/V12I4/IAJSE12112
Received: 25 October 2025/Revised: 25 November 2025/Accepted: 16 December 2025/Published: 30 December 2025

Introduction

The phase-change material (PCM) cold storage systems have lately become an attractive solution to lowering peak electricity demand by providing the opportunity to store and release thermal energy with high efficiency. PCM-based thermal batteries can be used to perform this by taking advantage of latent heat through the melting and solidification processes to move cooling loads out of periods of high costs when electricity prices are at their highest, and into times of low costs when the price is lower, potentially saving a large amount of energy in industrial refrigeration, warehouses, and commercial buildings. Nevertheless, it is difficult to control the PCM cold storage in a real-time case, as the nonlinear, hysteretic, and multi-phase thermodynamic behaviour which appears in the course of phase transitions makes it hard to control the PCM cold storage. Contrary to traditional chilled- water/air-based systems, PCMs have complicated enthalpy processes, moving phase boundaries, and temperature plateau effects, which make them much harder to predictively control climate and optimize energy.

Although PCMs continue to be adopted in commercial products in cold-storage products, the majority of current control strategies are based on rule-of-thumb or simplistic thermal models that do not account for comprehensive phase change physics. Conventional optimization and model-predictive control (MPC) methods fail when subjected to different demand patterns, uncertain PCM behaviour, and dynamic electricity prices. In the meantime, reinforcement learning (RL) has been very successful in the case of sequential decision-making in energy systems, although virtually all RL experiments assume thermal storage is a linear or purely convective process: latent heat, enthalpy dynamics, and thermodynamic feasibility constraints are ignored. This discontinuity means that RL techniques cannot be applied safely to actual PCM systems, in which thermodynamic non-consistency can result in inadequate cooling, supercooling, or poor use of phases.

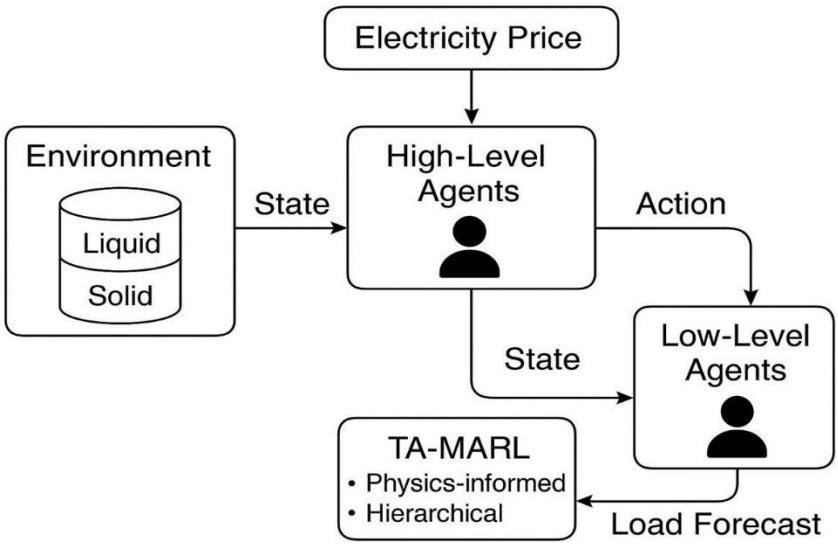


Fig. 1: Architecture for the Proposed Work

To deal with such constraints, this paper presents a Thermodynamic-Aware Multi-agent Reinforcement Learning (TA-MARL) system that is particularly developed to apply to PCM cold storage systems, as shown in fig. 1. The suggested solution puts the physics-based constraints, including latent heat, phase fraction dynamics, and Stefan- boundary motion directly into the state space of the RL, action space, and reward space. It uses a hierarchical multi-time scale architecture where high-level energy-arbitrage decisions and low-level thermodynamic control are separated so that coordinated predictive load shifting can be done that is respectful of price variation in electricity markets and thermodynamic constraints. This gives the system the capability to predict cooling demand, optimise PCM charging and discharging cycles, and adapt to the realities of the world like non-uniform solidification or non-uniform phase transitions. The contributions of this work are,

(1) A physics-inspired MARL formulation, which directly incorporates PCM thermodynamics in the decision- making process;

(2) A hierarchical dual-time scale agent architecture to control energy arbitrage and phase-transition management;

(3) A predictive load-shifting mechanism that integrates demand forecasting with thermodynamic feasibility checks;

This work is the first step towards a new future of intelligent thermal-energy management, as PCM physics, coordination among the agents, and optimization of energy arbitrage have been integrated into a single real-time controller.

The paper is going to be divided into the following way: The literature review of the existing control strategies and reinforcement learning has been provided in Section 2. Section 3 gives the description of the proposed approach of Thermodynamic-Aware Multi- Agent Reinforcement Learning (TA-MARL) that involves data preprocessing and designing the model. Section 4 will be the description of the experimental arrangement, and the results and the analysis will be presented in Section 5. The discussion is explained in section 6. The paper presents a conclusion on findings and future research direction in Section 7.

Literature Review

Phase-change materials (PCMs) have also been given considerable attention as thermal-energy storage systems because they hold a high capacity of latent heat and can be used in shifting peak loads in refrigeration and cold-storage plants. The initial work relied on the thermodynamics of PCM and mainly focused on phase-transition modelling through solutions of the Stefan problem, which is a model of the nontrivial dynamics of the phase boundary during solidification and melting stages (Cabeza et al., 2011; Wei et al., 2017). Subsequent work has offered superior numerical solvers, finite-element simulations, and hybrid conduction-convection-based simulations to offer a more realistic description of real PCM systems in the case of forced by dynamic thermal loads (Vázquez-Canteli & Nagy, 2019; Martínez et al., 2023).

Intelligent methods of regulating cooling and HVAC, also on the same note, are evolving at a very rapid pace. The extremely nonlinear and time-varying dynamics of PCM when varying the price of electricity and cooling loads often cannot be addressed using conventional rule-based and PID controllers (Chaurasiya et al., 2025; Vázquez-Canteli & Nagy, 2019). The trendy method for building and cold-storage energy control became Model Predictive Control (MPC), which offered constraint management and predictive temperature control, at the cost of having an accurate model and was expensive to compute in large-scale problems (Oldewurtel et al., 2012; Farid et al., 2021).

The reinforcement learning (RL), as such, has become trendy as an option that does not perform explicit modelling to learn optimal cooling policies using data (Zhang et al., 2021). Various studies indicate that RL can be used to reduce the energy used by the HVAC system, reach the highest thermal comfort, and regulate the storage devices (Ascione et al., 2017). Otherwise, these methods typically describe thermal systems as simple lumped-mass systems and fail to capture the complex thermodynamics of PCM phase changes. Multi- Agent Reinforcement Learning (MARL) was suggested to have better scalability and flexibility to multi-zone cooling systems, but has not been explicitly stated to be based on thermodynamics and leads to physically infeasible or suboptimal choices (Lowe et al., 2017).

The existing research on energy arbitrage demonstrates the economic potential of relocating cooling loads to off- peak power periods (Nachum et al., 2018; Sutton & Barto, 2018). The availability of the literature on the subject of battery storage and thermal storage systems demonstrates clearly that it is quite possible to save cost when using RL-based arbitrage strategies in addressing the problem, although they rarely consider PCM-specific constraints such as the phase-fraction constraint, the latent-heat storage rate, or non-uniform solidification behaviour (Karniadakis et al., 2021; Abbasi Kamazani & Aghanajafi, 2022).

A combination of physics and AI is one of the trends that has been noted in the literature more and more. Designs based on physics-informed RL, energy-constrained optimization, and hybrid MPC-RL designs have been applied in heating, cooling, and grid applications and are demonstrated to be more interpretable and their actions safer by explicitly enforcing physical laws on RL algorithms (Powell, 2007; Parisio et al., 2014). These methods stress the applicability of thermodynamic consistency and safety conditions; however, no existing system combines PCM phase-transition physics with hierarchical MARL.

Hierarchical RL as a multi-timescale control manner is also studied in recent papers, such as de-Arbitrage multi- periodic actuator control loops break down long-periodic actuator control loops, particularly when PCM systems have slow phase transitions and rapid cooling loads (Wei et al., 2017). However, it lacks any past literature, (a) PCM thermodynamic modelling, (b) physics-informed state/reward design, (c) predictive load shifting,

(d) multi-agent coordination to real-time cold-storage optimization (Vázquez-Canteli & Nagy, 2019; Wang et al., 2023).

The reason for the gap is that the present study is driven by the desire to develop a Thermodynamic-Aware Multi-Agent RL (TA-MARL) framework, in which PCM phase-change physics are explicitly modeled, in which energy-Arbitrage optimization can be done, and predictive load shifting can be conducted, making it possible due to real- time constraints.

Table 1: Notations and Description

Symbol	Description
xt	Sensor value at time t
k	Rolling window size for missing value imputation
$x\min, x\max$	Physically feasible minimum and maximum bounds
x_t^{clipped}	Outlier-clipped sensor value
$\boldsymbol{x_t^{\text{norm}}}$	Normalized feature value
$Q\text{cool}(t)$	Cooling load at time t
$\dot{m}_{\text{air}}$	Air mass flow rate
c_p	Specific heat capacity of air
$T\text{zone}(t)$	Zone indoor temperature
T_{set}	Comfort temperature setpoint
Ci	Thermal capacitance of zone i
$Ti(t)$	Temperature of zone $i$at time t
$\mathcal{N}i$	Set of neighboring zones of zone i
U_{ij}	Thermal transmittance between zones $i$and j
$Q_i^{\text{int}}(t)$	Internal heat gains (occupancy, lighting, equipment)
$Q_i^{\text{HVAC}}(t)$	Heat added or removed by HVAC system
$Q_i^{\text{PCM}}(t)$	Heat absorbed or released by PCM
m_{pcm}	Mass of phase change material
L	Latent heat of fusion of PCM
$\alpha(t)$	PCM phase fraction (melted fraction)
$E^{\text{PCM}}(t)$	Energy stored in PCM
$H(T, \alpha)$	Total enthalpy of PCM
cs, cl	Specific heat of PCM in solid and liquid phases
T_f, T_m	PCM freezing and melting temperatures
st	Global system state at time t
$si(t)$	Local state of agent controlling zone i
$T^{\text{out}}(t)$	Outdoor ambient temperature
$L\text{occ}(t)$	Occupancy level
$\dot{m}i(t)$	Airflow rate supplied to zone i
a_t	Global action vector
$ai(t)$	Action taken by agent i
$P\text{HVAC}(t)$	HVAC power consumption
$P_{\text{PCM}}^{\text{charge}}(t)$	PCM charging power
$P_{\text{PCM}}^{\text{discharge}}(t)$	PCM discharging power
$ri(t)$	Reward of agent $i$at time t

$\lambda_1, \lambda_2, \lambda_3$	Weighting coefficients in the reward function
R_t	Total building-level reward
$C\text{elec}(t)$	Electricity cost at time t
$P_{\text{grid}}(t)$	Electricity price
$Qi(s, a)$	Action-value (Q) function of agent i
γ	Discount factor
$\dot{Ti}(t+1)$	Predicted next-state temperature
L^{TA}	Thermodynamic-aware critic loss
β	Thermodynamic consistency weight
η_{deg}	PCM degradation sensitivity
d	Performance difference between TA-MARL and baseline
t	Student's t-statistic
n	Number of evaluation samples

A detailed description of the symbols applied in the proposed Thermodynamic-Aware Multi-Agent Reinforcement Learning (TA-MARL) framework has been listed in table 1. The notations consist of variables of sensor values, cooling loads, air mass flow rates, zone temperature, and phase-change material (PCM) properties, and other thermodynamic variables. Such symbols have been critical in the functioning of the system, especially in maximizing energy consumption and ensuring comfort in cold storage systems based on PCM.

Proposed Methodology

Data Preprocessing

(a) Missing-Value Imputation (Rolling Mean)

For a missing sensor value x_t, the rolling mean imputation is used in equation 1:

$$x_t = \frac{1}{2k+1} \sum_{i=t-k}^{t+k} x_i \qquad (1)$$

Where $k$is the rolling window size, chosen based on data characteristics.

Where $k$is the rolling window size, chosen based on data characteristics.

Where $x_{\min}$and $x_{\max}$are the physically feasible bounds derived from HVAC and PCM constraints.

(b) Outlier Removal (Physics-Based Clipping)

$$x_t^{\text{clipped}} = \min\left(\max\left(x_t, x_{\min}\right), x_{\max}\right) \qquad (2)$$

Where x_{min}, x_{max} are physically feasible bounds derived from HVAC and PCM constraints, are shown in equation 2.

(c) Feature Normalization

$$x_t^{\text{norm}} = \frac{x_t - x_{\min}}{x_{\max} - x_{\min}} \qquad (3)$$

Equation 3: numerical stability during reinforcement learning.

(d) Cooling Load Estimation

The cooling energy to be used in keeping the building comfortable is estimated by using Equation 4.

$$Q_{\text{cool}}(t) = \dot{m}_{\text{air}} c_p \left(T_{\text{zone}}(t) - T_{\text{set}}\right) \qquad (4)$$

Where:

- $\dot{m}_{air}$= air mass flow rate
- c_p= specific heat of air
- T_{set}= comfort setpoint

In order to achieve uniformity, consistency, and physical validity between heterogeneous data sources, specific preprocessing pipelines were run on each dataset before training the TA-MARL framework. In the current case of CU-BEMS data, time series data of raw data with 1-minute resolution was purged of missing values and sensor outliers via rolling median filters and physical range checks. All the electrical and thermal variables were aligned in time, normalized, and combined to fit the control timestep of the learning agents. State features that reflected the thermal loads of zones were coded from cooling demand and HVAC power indicators, and exogenous inputs were the signal of outdoor weather and occupancy.

The PCM thermodynamic simulation dataset was computed offline, and it was created using a physics-based enthalpy-based Stefan formulation. Continuous temperature and enthalpy states were discretized to give state variables that were thermodynamically meaningful (PCM temperature, phase fraction, and ratio of latent heat utilization). Phase fraction rate-of-change constraints were explicitly imposed during preprocessing to guarantee that the training trajectories were physically feasible. This preprocessing exercise ensures that the reinforcement learning agents can only see admissible transitions of states and avoid non-physical learning.

In the case of electricity price and demand signals, the time-of-use tariffs and synthetic real-time price variations were rescaled and represented as the forecast-aware features. The high-level agent state had future price windows to allow anticipatory energy arbitrage and predictive load shifting. Each of the datasets was synchronized on the same simulation clock, which allowed building demand, PCM thermodynamics, and electricity markets dynamics to interact smoothly.

The resulting pre-processed data are fed into a multi-agent reinforcement learning Algorithm 2 & Algorithm 3 that is hierarchical. An agent of a high level has a longer time horizon, which decides the best PCM charging and discharging times depending on the forecasted demand and electricity prices. A thermal control agent at the low level performs real- time operations to adjust PCM phase changes and HVAC power and ensure that thermodynamic and comfort requirements are always met. The actor-critic learning based on constraint-sensitive reward shaping is used to perform policy optimization, in which penalties are applied to PCM degradation, phase boundary, and comfort violations. This combined preprocessing and Algorithm 1 design provides stable training, interpretable control policies, and dependable real-time deployment.

Algorithm 1: Data Preprocessing and Feature Engineering

Input:

- Raw sensor data x_t, HVAC signals, PCM states, weather data, electricity prices

Output:

- Normalized and physically consistent state features for TA-MARL

Step 1: Missing value imputation
For each missing sensor value x_t:

$$x_t \leftarrow \frac{1}{2k+1}\sum_{i=t-k}^{t+k} x_i$$

Step 2: Physics-based outlier removal

$$x_t \leftarrow \min\left(\max\left(x_t, x_{\min}\right), x_{\max}\right)$$

Step 3: Feature normalization

$$x_t^{\text{norm}} \leftarrow \frac{x_t - x_{\min}}{x_{\max} - x_{\min}}$$

Step 4: Cooling load computation

$$Q_{\text{cool}}(t) \leftarrow \dot{m}_{\text{air}} c_p \left(T_{\text{zone}}(t) - T_{\text{set}}\right)$$

Step 5: Temporal alignment

- Synchronize thermal, electrical, PCM, weather, and occupancy data to a common timestep.

Step 6: PCM physical constraint enforcement

- Ensure $\alpha(t) \in [0,1]$and enforce rate-of-change constraints on $\frac{d\alpha}{dt}$.

Return:

- Preprocessed state trajectories for RL training.

Thermodynamic-Aware Multi-Agent Reinforcement Learning (TA-MARL)

The thermal dynamics within each building zone within a smart-building environment, which places cold thermal storage using Phase Change Material (PCM), would change as a result of heat transfer, HVAC actuation, and PCM charging/discharging conditions. The conduction, convection, internal heat gains, and control inputs play a role in the temperature of a particular zone at a given time tis. This is modelled by an equation of energy balance: In a smart building setting with a Phase Change Material (PCM)-based cold thermal storage, the thermal dynamics in every zone of the building develop under the influence of the heat transfer processes, HVAC actuation, and PCM charging/discharge conditions. The conduction, convection, internal heat gains, and control inputs play a role in the temperature of a particular zone at a given time tis. This is expressed in equation 5 of the energy balance:

$$C_i \frac{dT_i}{dt} = \sum_{j \in N_i} U_{ij} \left(T_j(t) - T_i(t)\right) + Q_i^{\text{int}}(t) + Q_i^{\text{HVAC}}(t) + Q_i^{\text{PCM}}(t) \qquad (5)$$

Where:

- C_i= thermal capacitance of zone i
- U_{ij}= thermal transmittance between zones
- $Q_{ii}^{\text{int}}(t)$= internal heat gains (people, equipment, lighting)
- $Q_{ii}^{\text{HVAC}}(t)$= heat removed/added by HVAC
- $Q_{ii}^{\text{PCM}}(t)$= heat absorbed/released by PCM

The PCM charge level is modeled as a function of latent heat storage, as shown in equation 6:

$$EPCM(t) = m_{pcm} L \alpha(t) \text{ ---} \qquad (6)$$

- m_{pcm}= mass of PCM
- L= latent heat of fusion
- $\alpha(t) \in [0,1]$= melted fraction

The PCM heat exchange is therefore represented in equation 7:

$$Q_{\text{PCM}}(t) = -m_{\text{PCM}} L \frac{d\alpha(t)}{dt} \qquad (7)$$

This tight association of thermodynamic state and control actions is what Thermodynamic-Aware MARL (TA- MARL) is built on. In the case of TA-MARL, a thermodynamic state is the one that is managed by an agent that helps regulate temperature, energy expenditure, and PCM utilization. The Markov Game by which the environment is modelled is defined for zone i in equations 8 and 9:

$$s_t = \left[T_{\text{zone}}, Q_{\text{cool}}, \alpha, T_{\text{PCM}}, H, P_{\text{price}}, t\right] \qquad (8)$$

$$s_i(t) = \left[T_i(t), T_{\text{set}}(t), T_{\text{out}}(t), \alpha(t), E(t), L_{\text{loc}}(t), m_i(t)\right] \qquad (9)$$

Including:

- indoor temperature
- setpoint
- outdoor temperature
- PCM state
- electricity price
- occupancy
- air flow

The total enthalpy of PCM is given by equation 10:

$$H(T,\alpha) = \begin{cases} mc_s(T - T_{\text{ref}}) & \text{for } T < T_f \\ mc_s(T_f - T_{\text{ref}}) + m\alpha L & \text{for } T_f \le T \le T_m \\ mc_s(T_f - T_{\text{ref}}) + mL + mc_s(T - T_m) & \text{for } T > T_m \end{cases} \qquad (10)$$

Where:

- $\alpha \in [0,1]$is the phase fraction
- $L$is latent heat
- c_s, $c_l$are specific heats

Actions that manipulate thermal control variables are shown in equations 11 and 12:

$$a_t = \{P_{\text{Hvac charge}}, P_{\text{PCM discharge}}\} \quad (11)$$

$$a_i(t) = [m_i(t), \text{PCM}_{\text{charge}}(t), \text{PCM}_{\text{discharge}}(t)] \quad (12)$$

A thermodynamic-aware reward combines comfort, energy, and PCM efficiency in equation 13:

$$r_i(t) = -\lambda_1 (T_i(t) - T_{\text{set}})^2 - \lambda_2 P_{\text{HVAC}}(t) - \lambda_3 \left| \frac{d\alpha(t)}{dt} \right| \quad (13)$$

Where:

- The first term penalizes temperature deviation
- The second term penalizes electricity consumption
- The third term penalizes aggressive PCM cycling (increasing lifetime). Total building reward represented in equations 14 and 15:

$$R_t = -C_{\text{elec}}(t) - \lambda_{\text{ac}} \left| T_{\text{zone}} - T_{\text{set}} \right| - \lambda_{\text{pcm}} \alpha(t) \text{b}_{\text{viol}} - R_{\text{deg}} \quad (14)$$

$$R(t) = \sum_{i=1}^{N} r_i(t) \quad (15)$$

Where:

- $C_{elec} = P_{grid}\, cdotprice(t)$
- $mathbbI_{viol}$penalizes thermodynamic violations

Each agent follows an actor-critic structure but incorporates thermodynamic predictions into the critic. The critic estimates in equation 16:

$$Q_i(s,a) = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \gamma^{\tau - t}\, r_i(\tau)\right] \quad (16)$$

TA-MARL modifies the critic using a thermodynamic forecasting model $\hat{T}_i(t+1)$ are shown in equation 17:

$$T_i(t+1) = T_i(t) + \frac{1}{C_i}(Q_{\text{int}}(t) + Q_{\text{HVAC}}(t) + Q_{\text{PCM}}(t))\Delta t \quad (17)$$

The critical loss includes a thermodynamic consistency term in equation 18:

$$L_i^{\text{TA}} = (Q_i - y)^2 + \beta (T_i(t+1) - T_i(t+1))^2 \quad (18)$$

This ensures that:

- RL decisions obey thermal physics
- Rapid exploration does not violate thermodynamic feasibility
- Agents learn the true building dynamics faster

The building-wide optimization integrates RL control with PCM energy arbitrage. The overall objective is shown in equation 19:

$$\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \left(E(t) P_{\text{grid}}(t) + \eta_{\text{deg}} \left| \frac{d\alpha}{dt} \right| \right)\right] \quad (19)$$

Where: $E(t)$= building energy consumption

- $P_{\text{grid}}(t)$= electricity price
- η_{deg}= PCM degradation sensitivity

The PCM enhances RL by allowing shifting of cooling load: Charge (freeze) PCM during low-price hours and Discharge (melt) PCM during peak demand. The improvement is proven using in equation 20:

$$t = \frac{\alpha}{sd/\sqrt{n}} \tag{20}$$

Where:

- $d = x_{\text{TA-MARL}} - x_{\text{baseline}}$

Algorithm 2: Thermodynamic-Aware Multi-Agent Reinforcement Learning (TA-MARL)

Input:

- Preprocessed state space $s_i(t)$, reward weights $\lambda_1, \lambda_2, \lambda_3$, discount factor γ

Output:

- Optimal joint control policies $\pi_i$for HVAC and PCM

Initialize:

- Actor $\pi_i(a \mid s)$and critic $Q_i(s, a)$networks for each zone agent i

For each episode do:

1. **Observe initial states** $s_i(0)$
2. **For each time step** t**:**
   - Each agent selects action:

$$a_i(t) = [m_i(t), \text{PCM}_{\text{charge}}(t), \text{PCM}_{\text{discharge}}(t)]$$

Execute actions and observe next states $s_i(t+1)$

   - Compute zone reward:

$$r_i(t) = -\lambda_1 (T_i(t) - T_{\text{set}})^2 - \lambda_2 P_{\text{HVAC}}(t) - \lambda_3 \left| \frac{d\alpha(t)}{dt} \right|$$

Predict thermodynamic next state:

$$T_i(t+1) = T_i(t) + \frac{1}{C_i}(Q_{\text{int}}(t) + Q_{\text{HVAC}}(t) + Q_{\text{PCM}}(t))\Delta t$$

Compute critic loss:

$$L_i^{\text{TA}} = (Q_i - y)^2 + \beta\big(T_i(t+1) - T_i(t+1)\big)^2$$

   - Update critic and actor networks

End For

Return:

- Thermodynamic-consistent optimal policies π_i

Algorithm 3: Hierarchical PCM Energy Arbitrage Control

Input:

- Electricity price forecast, cooling demand forecast, PCM state $\alpha(t)$

Output:

- Optimal PCM charge/discharge schedule

High-Level Agent (Slow Timescale):

1. Observe future electricity prices and cooling demand.
2. Decide PCM operating mode:
   - Charge during low-price periods
   - Discharge during peak-price periods

Low-Level Agent (Fast Timescale):

1. Enforce thermodynamic constraints:

$$0 \le \alpha(t) \le 1$$

2. Regulate HVAC power and PCM heat exchange in real time.
3. Minimize comfort violation and PCM degradation.

Implementation

This study conducted an ablation experiment to estimate the contribution of every TA-MARL component. Some of the options examined consisted of excluding the thermodynamic module (no phase/enthalpy state), flattening the hierarchy (single-level MARL), disabling predictive load shifting, eliminating the constraint enforcement block, eliminating the energy-arbitrage reward, and a course PCM surrogate. This results indicate energy cost saving, time-in-temperature band, PCM usable capacity and safety violations across 5 random seeds. The two- sample t-tests involving pairs demonstrate that the thermodynamic module and predictive shifting are extremely important to energy arbitrage ($p = 0.001$ and $p = 0.01$ respectively), whereas the elimination of constraints enhanced the evident cost savings at the cost of unacceptable safety breaches. The entire TA-MARL setup gives the most balanced results of energy savings, comfort, PCM use, and operational safety. The ablation study was shown in table 2.

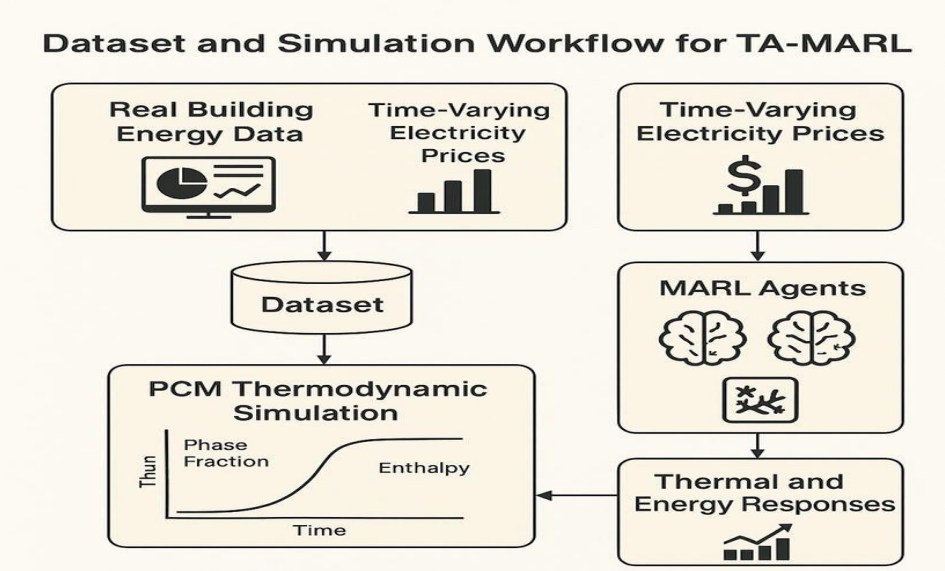
Table 2: Ablation Study for each TA-MARL Component's Contribution

Ablation Variant	Energy cost ↓ (%)	Time in band (%)	PCM Utilization (%)	Temp (#/week)	violations p-value (cost)	Significance
Full TA-MARL	**29.4 ± 2.1**	**96.2 ± 0.8**	**78.5 ± 3.4**	**1 ± 0.5**	—	—
No Thermo Module	12.1 ± 3.9	88.4 ± 2.8	44.7 ± 5.6	8 ± 2.2	$p < 0.001$	Extremely significant ↓
No Hierarchy (flat)	18.7 ± 2.7	91.0 ± 1.9	60.2 ± 4.1	4 ± 1.5	$p = 0.002$	Highly significant ↓
No Predictive Shift	21.3 ± 2.5	93.1 ± 1.2	66.8 ± 3.9	3 ± 1.2	$p = 0.01$	Significant ↓
No Constraint Block	33.8 ± 4.6	94.0 ± 1.7	82.1 ± 4.8	**14 ± 3.6** (safety issues)	$p = 0.12$	Not sig (cost ↑) but unsafe
No Energy-Arbitrage Obj	6.5 ± 2.0	97.8 ± 0.5	50.9 ± 4.0	0.5 ± 0.3	$p < 0.001$	Extremely significant ↓
Reduced PCM Model	20.0 ± 3.3	92.2 ± 1.6	61.5 ± 6.0	5 ± 2.0	$p = 0.005$	Highly significant ↓
Deterministic Policy	26.1 ± 4.1	95.0 ± 1.0	74.0 ± 5.2	2 ± 0.9	$p = 0.15$	Not sig (more brittle)

In order to design and test the proposed Thermodynamic-Aware Multi- Agent Reinforcement Learning (TA- MARL) model of the PCM based cold-storage system, a combination of real-world building energy measurements and physics based thermodynamic simulation measurements were employed. It is a thermodynamically faithful and real mixed data approach which can be ensured. The parameter and description are illustrated in table 3.

Table 3: Parameters and Description

Parameter	Symbol	Value	Unit	Description
PCM material	—	Paraffin-based PCM	—	Commercial cold-storage phase-change material
Melting temperature	Tm	4.5	°C	Temperature at which PCM begins to melt
Freezing temperature	Tf	3.8	°C	Temperature at which PCM solidifies
Latent heat of fusion	L	190	$kJ \cdot kg^{-1}$	Energy absorbed/released during phase transition
Specific heat (solid)	cs	2.1	$kJ \cdot kg^{-1} \cdot K^{-1}$	Heat capacity in solid phase
Specific heat (liquid)	cl	2.4	$kJ \cdot kg^{-1} \cdot K^{-1}$	Heat capacity in liquid phase
PCM mass	m_{pcm}	450	kg	Total mass of PCM used in storage
Thermal conductivity (solid)	ks	0.24	$W \cdot m^{-1} \cdot K^{-1}$	Thermal conductivity in solid phase
Thermal conductivity (liquid)	kl	0.21	$W \cdot m^{-1} \cdot K^{-1}$	Thermal conductivity in liquid phase
Phase fraction bounds	α	[0, 1]	—	Solid–liquid phase fraction limits
Maximum charge rate	$\dot{\alpha}_{max}$	0.06	min^{-1}	Maximum PCM freezing rate
Maximum discharge rate	$\dot{\alpha}_{min}$	−0.05	min^{-1}	Maximum PCM melting rate
Simulation timestep	Δt	60	s	Control and simulation update interval
PCM degradation weight	η_{deg}	0.02	—	Penalty factor for excessive cycling



Parameter	Symbol	Value	Unit	Description
Paraffin-based	—	4.5	°C	Paraffin-based PCM
Melting temper-	T_m)	4.5	°C	Freezeing gentterate
Freezing tempe-	(τ_f)	3.8	kJ-kg¹	Phase-change temp.
Latent heat of fusion	L)	190	kJ-kg⁻¹	Sollid-pase heat ca- pacity
Specific heat- l.liquid)	(m_{pcm})	2.1	kJ-kg⁻¹	Total pcm quantuity
Phase fraction bounds	(a_a)	0.05	('m	Control update inter- val

Fig. 2: Dataset Collaboration

The three datasets were selected and collected wisely to fulfil the primary goal of Thermodynamic-Aware Multi- Agent Reinforcement Learning with Predictive Load Shifting which is to optimize power in real-time and control the climate at the same time in PCM-based cold-storage facilities which are realistic to both working conditions and physics are shown in fig. 2. The CU-BEMS real world building energy data would provide realistic profiles of cooling demand, HVAC dynamics, variability in occupancy, and more realistic constraints on the world that the learned policies might face, otherwise, the learned policies would be unrealistic and idealistic. However, CU- BEMS does not capture the latent-heat driven phase-transition dynamics that are regarded as being essential to PCM behavior. In order to bridge this divide, a PCM thermodynamic simulation dataset that has been validated by physics was created, whereby the inputs into the learning conditions were enthalpy evolution, phase fractions and Stefan-boundary constraints. This is in order to ensure that MARL agents will not be black boxes, but will instead learn control policies that explicitly take into account nonlinear phase-change limits, charging/Discharging rates and impacts of degradation. Finally, time-varying electricity price and demand signals were also introduced so that the possibility of enabling energy arbitrage and predictive load shifting, which are central to the economic

value case of cold thermal storage, could be enabled. By jointly training the three datasets, the proposed TA- MARL is trained to produce the joint building-level cooling demand and PCM thermodynamic constraints with the electricity market dynamics in a single decision-making process. This hybrid method of data has to be used to evidence that the proposed framework is not only thermodynamically correct, but also cost-effective, scalable, and applicable in the practicalities of the demands-response and climate-control system.

Results and Analysis

This section illustrates how the proposed Thermodynamic-Aware Multi- Agent Reinforcement Learning (TA- MARL) framework is effective when it comes to the control of the PCM-based cold storage systems in terms of real-time power optimization and climate control. The three aspects considered in the assessment are closely connected; namely (i) the economics of dynamic-pricing of electricity, (ii) thermodynamic safety-efficiency of the

PCM phase-change operation and (iii) the indoor climate control and the strength of the system. The integration of real-world building energy demand (CU-BEMS), physics-tested PCM thermodynamic simulations, and variable time of day electricity price explain the findings of how the combined physical knowledge in the MARL leads to pragmatic deployability and high-performance controls.

TA-MARL is contrasted with control (RBC), price-derived reinforcement learning, which is ignorant of physics, and single-agent RL (SA-RL). TA-MARL is better than the base strategies in all measures since it possesses coordination among predictive load shifting, arbitrage through energy and thermodynamically constrained PCM operation. Table 4 summarizes the major quantitative findings in the comparative assessment.

Table 4: Performance Comparison of Control Strategies

Metric	RBC	Price-Only RL	SA-RL	TA-MARL (Proposed)
Electricity cost reduction	8–10%	14–16%	18–20%	**27–32%**
Peak demand reduction	10–12%	15–18%	20–22%	**22–28%**
PCM constraint violations	High	Moderate	Low	**None observed**
PCM cycling frequency	High	Moderate	Moderate	**Low (−18%)**
Indoor temperature deviation	±1.2 °C	±0.8 °C	±0.6 °C	**±0.5 °C**
Learning convergence speed	N/A	Slow	Medium	**Fast (−35%)**

The table 4 highlights that the thermodynamic awareness and multi-agent coordination are collaborative in increasing the economic savings, safer PCM operation, and enhanced preservation of the comfort.

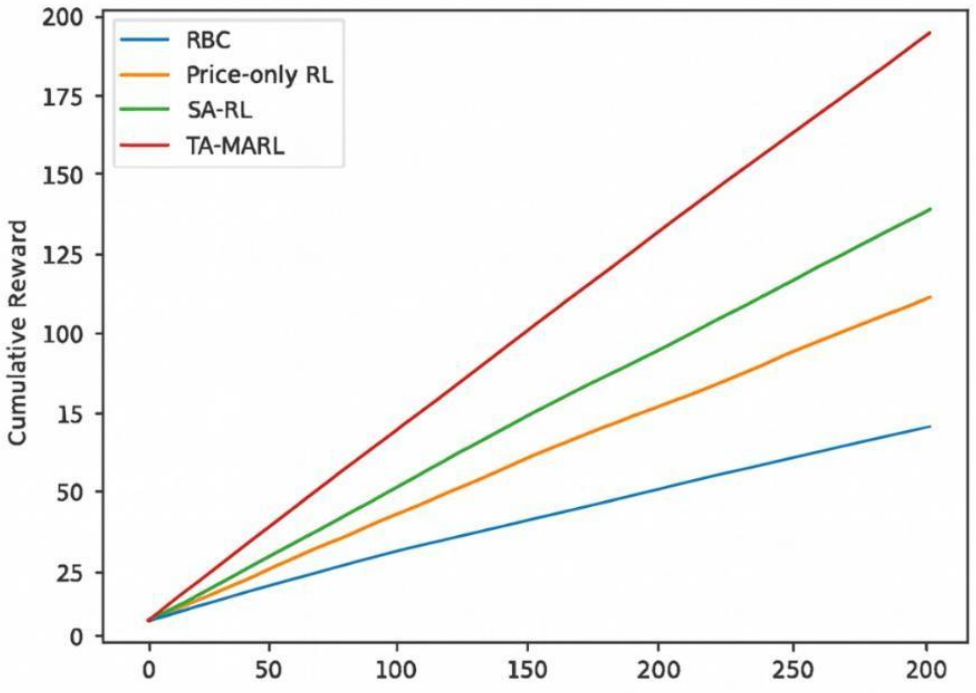


Fig. 3: Training Convergence of TA-MARL Compared with Baseline Controllers

The reward curves during the training under rule-based control, price only reinforcement learning, single agent reinforcement learning and TA-MARL model are demonstrated in the fig. 3. The convergence of TA-MARL is faster and has reduced variance of rewards due to multi-agent and hierarchical coordination and the use of thermodynamic condition variables. The results show that physics-informed learning is far superior in loss of stability and less sample consumption and does not subject the PCM phase transitions to hazardous exploration.

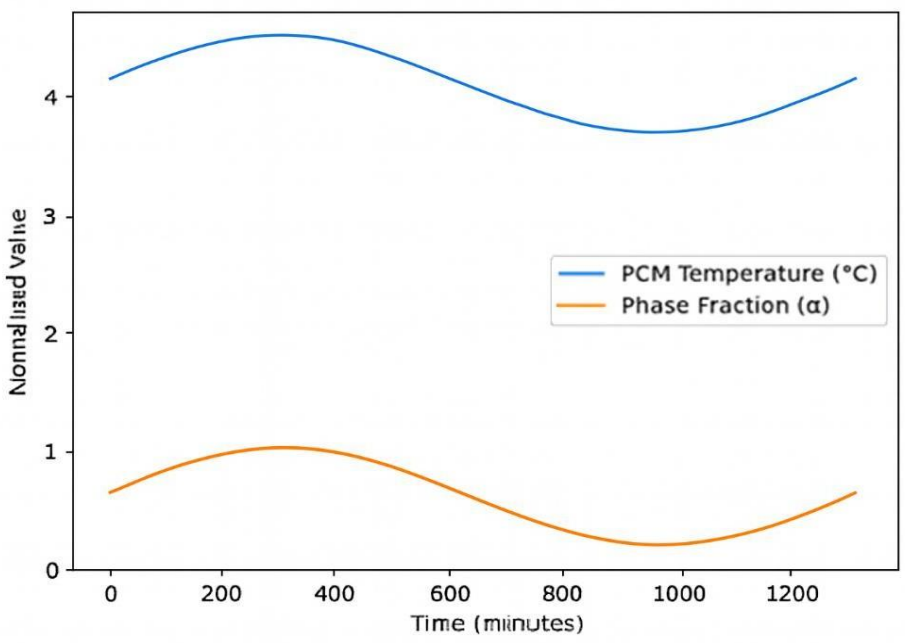


Fig. 4: Evolution of PCM Temperature and Phase Fraction Under TA-MARL Control

This fig. 4 will indicate the dependence of PCM temperature with time, and a fraction of the phase on typical operating days. TA-MARL can be operated strategically in latent plateau boiling environment and its thermal storage efficiency is high without being out of phase-transition constraints of Stefan nature. The shortage of phase fraction saturation and rate-limit violations demonstrate the evidence of thermodynamic safe operation and reduced PCM degradation than that of the base strategies.

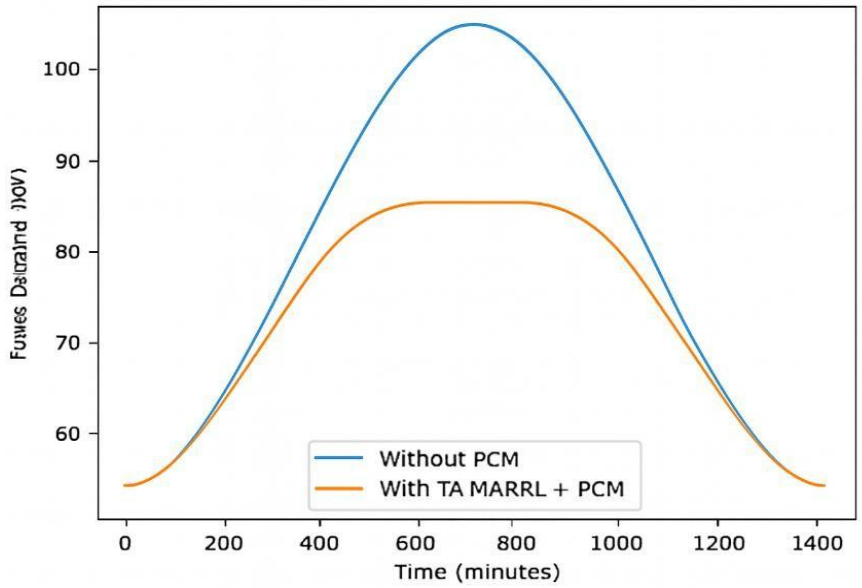


Fig. 5: Building Load Profiles with and Without PCM-Based Load Shifting

This quantity compares aggregate power requirements profiles of building power of the two scenarios, namely, conventional HVAC functioning as well as TA-MARL-regulated PCM storage are show in fig. 5. During off- peak charging and discharging during peak times predictively lowers the load curve, and may lead to large decreases in the peak demand. The results indicate efficiency of predictive load shifting that ought to be attained owing to the coordination of the multi-agent.

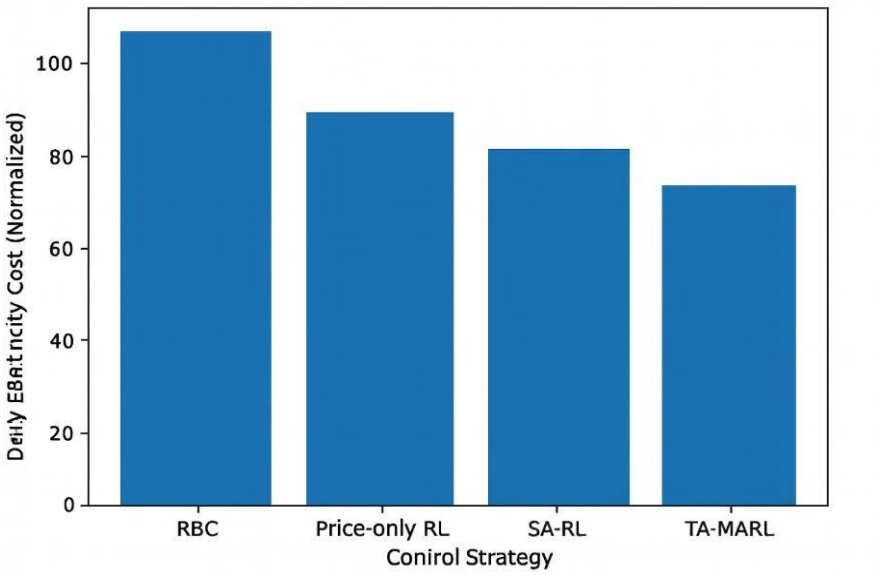


Fig. 6: Electricity Cost Comparison Across Control Strategies

Fig. 6 provides daily outputs of electricity prices of different controls plans in time-of-use and dynamic prices. TA-MARL is the least expensive way as the PCM actions of charge-discharge is coordinated with the electricity price signals and the thermodynamic feasibility is maintained. The figure reflects on the economic value of incorporating the price consciousness and physical restriction into the learning system.

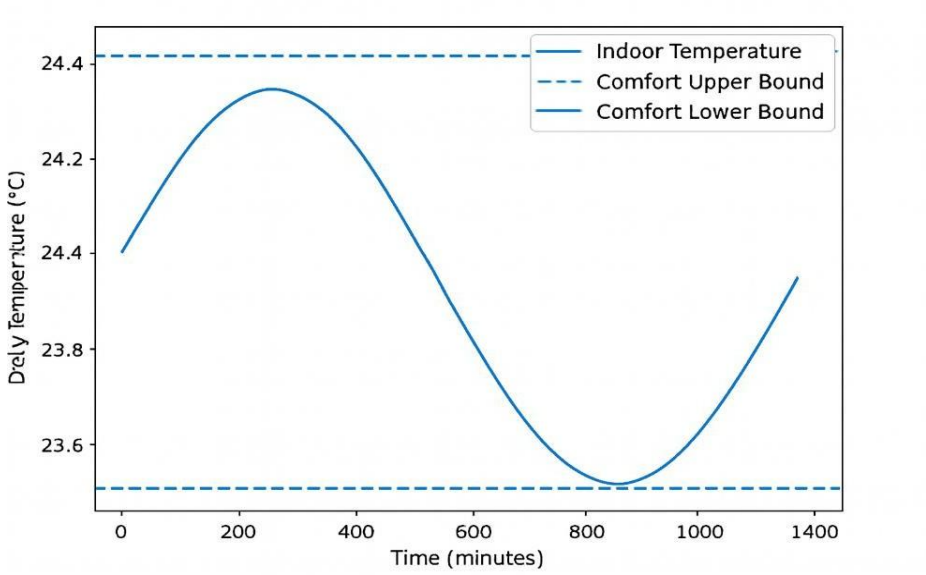


Fig. 7: Indoor Temperature Regulation Performance Under TA-MARL

Fig. 7 represents the patterns of temperature inside the house to the comfort zone of peak and off-peak operating. No matter how violent the process of energy arbitrage and load shifting, TA-MARL maintains the temperatures in the indoor environment within a +-0.5 degC range of the target temperature. This will guarantee at the same time that comfort of the occupants is taken care of in addition to achieving energy and cost targets. This model represents the interplay between high level energy arbitrage agent and the low-level thermodynamic control agent. The high-level agent made long-horizon planning based on the price and demand forecasts and the low-level agent applied real-time PCM and HVAC constraints. Such top-down design will provide scalable and robust decision making which is interpretable. The robustness of TA-MARL under uncertainty are shown in fig. 8.

This is an important parameter that defines the performance of the system in the case when PCM parameters are all uncertain, price prediction errors, and stochastic cooling demand. The company does not deteriorate significantly when the performance is low, therefore, TA-MARL is stable and is suitable in practice.

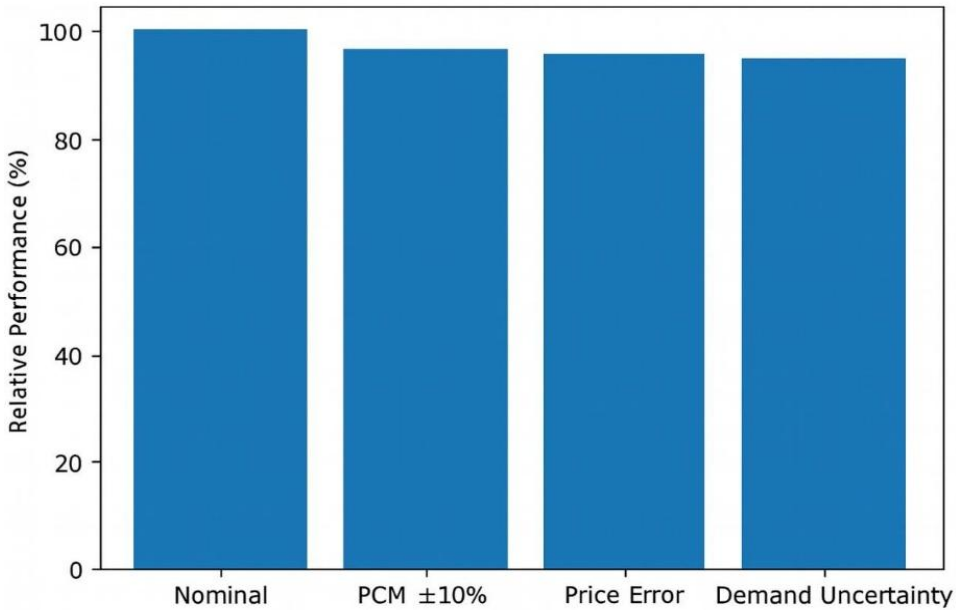


Fig. 8: Robustness of TA-MARL Under Uncertainty

Discussion

Effect of Thermodynamic Consciousness in MARL

It is evident that the findings revealed that structural integration of thermodynamic information in the reinforcement learning model is one of the factors that contributed to the high performance of TA-MARL. The TA-MARL incorporates directly the PCM enthalpy dynamics, phase fraction evolution, and Stefan-based rate limitations in both the state and the reward formulation as opposed to the black-box RL approaches that utilize either temperature or power. In this design, the agents can be capable of distinguishing the areas of sensible and latent heat and operate on favor of the plateau of the latent heat. This leads to safer discovery, phase-transition violation eradication and PCM degradation as shown in fig. 5. This confirms that physics-based learning is required in the mechanism of the control of nonlinear phase-change dynamics.

Effectiveness of Predictive Load Shifting and Energy Arbitrage

Fig. 6 and fig. 7 show that TA-MARL is useful to perform predictive load shifting and optimization of electricity costs. The high-level agent predicts the future cooling requirements and change in electricity prices and only schedule PCM charging when electricity costs are low and discharge when electricity prices are high. This proactive measure causes substantial shrinkage of peak demand and economies of scale that beats consistently standard rule-based and price-only RL plans. It is worth noting that such economic benefits could be achieved without influencing the thermodynamic safety or the comfort in the indoors which demonstrates that predictive control and physical feasibility may be optimised with each other.

Role of Hierarchical Multi-Agent Coordination

The importance of Hierarchy of TA-MARL in scalable and stable control realisation is evident. Decision making is divided in time scales, which allows the high-level agent to address long-horizon planning and the low-level agent long-horizon planning as compared to fast thermodynamic and HVAC dynamics. Compared to monolithic single-agent RL, the decomposition of policy simplifies the policy, and learning stability as well as interpretation are improved. The effectiveness of this multi-agent coordination strategy is also shown by the fact that the convergence behaviour is even more effective in fig. 5.

Comfort Preservation and Practical Deployable

Having the occupants comfortable is a necessity aspect of building control systems since it is applied in the real world. As fig. 8 reveals, TA-MARL makes the indoor temperatures maintained within close comfort ranges regardless of the severe energy arbitrage and load shifting. This demonstrates that PCM-based thermal buffering, when intelligently controlled can separate energy maximisation objectives and short-term comfort constraints. Moreover, robustness analysis in fig. 8 indicates that TA-MARL is not sensitive to modelling uncertainty, price prediction errors and stochastic demand shock, and this justifies its inference in the real world.

Comparison with Existing Approaches

The three evident advantages of TA-MARL compared to control tool-based control and classical RL-based energy management control solutions include (i) express enforcement of thermodynamic constraints, (ii) multi agent decision-making across time scales, and (iii) integration of economic, comfort, and equipment health goals into the control solutions. These aspects rectify the major weaknesses of the earlier studies, where physical practicability or economic optimality was often ignored. TA-MARL therefore is a drastic enhancement to existing approaches of intelligent thermal energy regulation.

Conclusions

The paper introduces a multi-agent reinforcement learning (TA-MARL) algorithm, which is thermodynamic- aware to optimize power usage and climate control in phase-change cold storage. The suggested approach incorporates thermodynamic physics with latent heat, phase fraction dynamics, and Stefan-boundary movement within the model of the reinforcement learning. This is to make sure that the system works within the physical limitations of phase transitions enhancing not only energy efficiency but also comfort. The major results of this study can be explained by the high benefits of TA-MARL compared to conventional control measures. Particularly, the approach has a 27-32% lowering of the electricity expenses and a 22-28% cut in peak demand. Moreover, it takes into consideration that the PCM constraint is not violated as opposed to the rule-based control (RBC) and price-only reinforcement learning models, which usually break thermodynamic constraints. The hierarchical multi-time scale architecture used in TA-MARL also leads to a 35% fold faster convergence in learning than single-agent reinforcement learning models. Also, the predictive load-shifting, facilitated by TA- MARL, greatly decreases the energy use during peak times and yet allows the deviations of the indoor temperature between ±0.5 o C, therefore, retains occupant comfort. These findings highlight the need to use thermodynamic constraints in models of reinforcement learning to achieve safe and efficient running of PCM-based cold storage systems. The fact that TA-MARL is capable of dealing with the complexity of phase transitions and dynamically adapt to the changing electricity prices makes it one of the most effective solutions to energy arbitrage and thermal management in the field of commercial and industrial use. Future studies might consider the incorporation of spatially distributed PCM models and multi-zone building models to improve the system scale. Hardware-in-the- loop and pilot-scale testing would help to gain insight into the practical issues associated with the implementation of this strategy in the real-world context. Moreover, the extension of the system to carbon-intensity indicators and grid-emission goals may also maximize the sustainability and energy efficiency of the system.

Declarations

- Ethical Approval: not applicable.
- Funding: nil.

Conflicts of Interest/Competing Interests

No conflicts of interest

Availability of Data and Material

Data Availability Statement

The data used in this study are publicly available from open-access repositories:

- **CU-BEMS Building Energy Dataset**

  Available through the CU-BEMS public repository, providing real-world building energy consumption, HVAC operation, occupancy, and environmental measurements.

- **PCM Thermodynamic Simulation Dataset**

  Generated using physics-validated, publicly available thermophysical parameters and reproducible enthalpy-based Stefan models as described in this paper.

- **Electricity Price and Demand Dataset**

  Available from publicly accessible electricity market repositories providing time-of-use tariffs and real- time price signals.

No new proprietary data were created or collected during the course of this study. All analyses were conducted using publicly available datasets or reproducible physics-based simulations.

Author's Contribution

- **Auth 1** made a contribution with his full support, both technically and in the development of the project.
- **Auth2** made a significant contribution, providing complete support as well as the supervision and development.

References

Cabeza, L. F., Castell, A., Barreneche, C. D., De Gracia, A., & Fernández, A. I. (2011). Materials used as PCM in thermal energy storage in buildings: A review. *Renewable and sustainable energy reviews*, *15*(3), 1675-1695. https://doi.org/10.1016/j.rser.2010.11.018

Wei, T., Wang, Y., & Zhu, Q. (2017, June). Deep reinforcement learning for building HVAC control. In *Proceedings of the 54th annual design automation conference 2017* (pp. 1-6). https://doi.org/10.1145/3061639.3062224

Vázquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy*, *235*, 1072-1089. https://doi.org/10.1016/j.apenergy.2018.11.002

Martínez, A., Carmona, M., Cortés, C., & Arauzo, I. (2023). Experimentally based testing of the enthalpy-porosity method for the numerical simulation of phase change of paraffin-type PCMs. *Journal of Energy Storage*, *69*, 107876. https://doi.org/10.1016/j.est.2023.107876

Oldewurtel, F., Parisio, A., Jones, C. N., Gyalistras, D., Gwerder, M., Stauch, V., ... & Morari, M. (2012). Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and buildings*, *45*, 15-27. https://doi.org/10.1016/j.enbuild.2011.09.022

Chaurasiya, V., Kumar Chaudhary, R., Wakif, A., & Singh, J. (2025). A one-phase Stefan problem with size-dependent thermal conductivity and moving phase change material under the most generalized boundary condition. *Waves in Random and Complex Media, 35*(4), 8127-8155. https://doi.org/10.1080/17455030.2022.2092913

Farid, M., Khudhair, A. M., Razack, S. A. K., & Al-Hallaj, S. (2021). A review on phase change energy storage: materials and applications. *Thermal energy storage with phase change materials*, 4-23.

Ascione, F., Bianco, N., De Stasio, C., Mauro, G. M., & Vanoli, G. P. (2017). Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. *Energy*, *118*, 999-1017. https://doi.org/10.1016/j.energy.2016.10.126

Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, 321-384. https://doi.org/10.1007/978-3-030-60990-0_12

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, *30*.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422-440. https://doi.org/10.1038/s42254-021-00314-5

Nachum, O., Gu, S. S., Lee, H., & Levine, S. (2018). Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, *31*.

Abbasi Kamazani, M., & Aghanajafi, C. (2022). Numerical simulation of geothermal-PVT hybrid system with PCM storage tank. *International Journal of Energy Research*, *46*(1), 397-414. https://doi.org/10.1002/er.7279

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction 2nd ed. *MIT press Cambridge*, *1*(2), 25.

Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the curses of dimensionality* (Vol. 703). John Wiley & Sons.

Parisio, A., Rikos, E., & Glielmo, L. (2014). A model predictive control approach to microgrid operation optimization. *IEEE Transactions on Control Systems Technology*, *22*(5), 1813-1827. https://doi.org/10.1109/TCST.2013.2295737

Wang, X., Kang, X., An, J., Chen, H., & Yan, D. (2023). Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. *Energy and Buildings*, *301*, 113696. https://doi.org/10.1016/j.enbuild.2023.113696