Physics-Based Reinforcement Learning Framework for Energy Management of Airport Thermo-Electrical Microgrids

Pablo Verdugo-Rivadeneira, *Graduate Student Member, IEEE,* Mehrdad Pirnia, *Member, IEEE,* Claudio A. Cañizares, *Fellow, IEEE,* Gonzalo E. Constante-Flores, *Member, IEEE*

***Abstract*—With the adoption of electric planes to achieve aviation Net-Zero goals, the implementation of proper Energy Management Systems that can satisfy both the electric and thermal needs of airport buildings is paramount. However, the design of Energy Management Systems for thermo-electrical microgrids is challenging, particularly for model-based approaches, due to the complexities associated with detailed component modeling, uncertainties in renewable generation, and environmental conditions. Thus, this paper proposes a model-free Reinforcement Learning framework for the optimal operation of multi-zone airport hangar microgrids ensuring constraint satisfaction by incorporating physics-based dynamic constraints, and an implicit method to properly represent heat exchanges between zones. Simulation results, based on the operation of an actual Canadian airport microgrid with e-plane charging are benchmarked against a classical optimization-based approach, demonstrating that the proposed method achieves near optimal results. Furthermore, compared to conventional reward-based approaches, the proposed framework is shown to have significantly faster convergence and more stable training behavior.**

***Index Terms*—Airport microgrid, Energy Management System, multi-zone building thermal model, reinforcement learning.**

NOMENCLATURE

Sets and Indices

$\underline{\cdot}, \bar{\cdot}$	Lower and upper limits
$\{nc\}$	Rooms without active temperature control
$\{ng\}$	Non-ground-level rooms
$\{ph\}$	Planes' hangar room
$e \in \mathcal{B}$	Building envelope surfaces
$h \in H$	Heat pumps (HPs)
$i, j \in \mathcal{I}$	Rooms
$k \in T$	Time steps
$s_k \in \mathcal{S}$	System state at time step k

Parameters

Δt_k	Time interval between step k and step $k+1$ [h]
$\eta_n^{\text{ch/dch}}$	Battery charging/discharging efficiency
γ	Discount factor for future rewards
λ_1	BESS degradation penalty coefficient
λ_2	Temp. violations penalty coefficient
Vol_i	Volume for room i $[\text{m}^3]$
μ_i	HP thermal distribution factor for room i
ψ_h^{HP}	HP heating efficiency reduction coefficient
ρ^{air}	Air density [kg/m^3]
$\sigma_{e,i}$	Shadowing coefficient of surface e
$\tau_{e,i}$	Transmission coefficient of windows
θ^{gr}	Ground temperature [K]
$\theta_h^{\text{HP,out}}$	HP outlet temperature [K]
θ_k^{ext}	Ambient temperature [K]
$\xi_{e,i}$	Amount of windows of building surface e
A_i^{f}	Area of building floor $[\text{m}^2]$
$A_{e,i}$	Area of building surface $[\text{m}^2]$
A_{ij}^{w}	Area of internal wall between room i and j $[\text{m}^2]$
c^{air}	Air specific heat [kJ/kgK]
c_k^{gp}	Cost of grid power [\$/kWh]
$c_k^{\text{BESS/SOC}}$	BESS penalty for degradation/target SOC [\$]
c_k^{temp}	Cost for temperature violations [\$]
C_i^{in}	Heat capacity of indoor air in room i [kJ/K]
E_n	BESS capacity [kWh]
I_k	Solar irradiation [kW/m^2]
P^{bat}	BESS rated power [kW]
$P_k^{\text{d/ep}}$	Building/e-plane active power demand [kW]
P^{EH}	EH rated active power [kW]
P_h^{bat}	HP rated power [kW]
P_k^{PV}	PV active power [kW]
RC	BESS replacement cost [\$/kWh]
TOU_k	Time of Use [\$/kWh]
$U_{e,i}$	Transmittance of external walls [kW/(m^{2}K)]
U_{ij}^{w}	Transmittance of internal walls [kW/(m^{2}K)]
U_i^{f}	Transmittance of floor [kW/(m^{2}K)]
$v_{i,k}^{\text{rate}}$	Ventilation rate [1/h]

Variables

$\mathbf{b}_k$	RHS vector (prev. temps. + thermal inputs) [K]
$\mathbf{M}$	Conductive coupling matrix
Φ_k	Battery degradation function
ϕ	Trainable parameters of the DRL policy
π_ϕ	DRL-based EMS policy parameterized by ϕ
COP_k	HP Coefficient of Performance
$\theta_{i,k}^{\text{in}}$	Indoor temperature in room i [K]
a_k^{EH}	EH discrete control action (1=ON)
$a_{h,k}^{\text{HP}}$	HP continuous control actions [kW]
a_k^{bat}	BESS continuous control action [kW]
$J(\pi_\phi)$	Expected return of policy π_ϕ [\$]
P_k^{gp}	Active grid power [kW]
$P_{h,k}^{\text{HP}}$	Active power of HP h [kW]
$Q_{i,k}^{\text{eq/lg/oc}}$	Heat from equipment, lights, and occupants [kW]
$Q_{h,k}^{\text{HP}}$	Thermal power of HP h [kW]
$Q_{i,k}^{\text{env/gr/vent}}$	Heat exchange through building envelope, ground, and ventilation [kW]
$Q_{i,k}^{\text{ig/sun}}$	Internal heat gains/solar thermal power [kW]
$Q_{ij,k}^{\text{trf}}$	Thermal power exchanged between rooms [kW]
r_k	Reward function [\$]
SOC_k	BESS State of Charge (SOC) [kWh]

I. INTRODUCTION

IN 2022, the 184 member states of the International Civil Aviation Organization (ICAO) adopted a long-term global aspirational goal of achieving net-zero carbon emissions from international aviation by 2050 [1]. To meet this target, several alternatives are being explored for the energy supply of Electric Aircraft (EA) and airport buildings, including the use of alternative fuels such as hydrogen; the adoption of hybrid or EA; the electrification of ground transportation operations; the integration of Distributed Energy Resources (DERs); and the deployment of energy-efficient buildings with Thermo-Electrical (TE) energy systems.

Microgrids (MGs), which are clusters of uncontrolled loads and DERs, with the latter comprising Distributed Generation (DG), Energy Storage Systems (ESSs), Renewable Energy Sources (RESs), and controllable loads [2], have been considered as a means to reliably satisfy the TE needs of airports. However, due to the large number of associated components and the inherent modeling complexities, developing an adequate Energy Management System (EMS), capable of coordinating electrical and thermal subsystems to achieve reliable and economical MG operation, remains challenging.

Although research on airport MGs is on the rise, the integration of airport TE-MG EMSs remains largely unexplored. With the exception of [3], which investigates the TE energy management in an airport building, using electricity, heating, and cooling demand profiles based on Monte Carlo simulations, and [4], where an optimization-based Model Predictive Control (MPC) approach is used to determine the optimal operation of a multi-zone airport hangar MG, most existing studies focus on other aspects of airport MG management. For instance, airport MG planning to accommodate Electric Vehicle (EV) parking lots and EA is addressed in [5], while [6] studies the planning of ground transportation and Battery Energy Storage System (BESS) capacity. In [7], the authors propose flight scheduling and hybrid EA charging optimization schedules, and [8] evaluates the impact of EV and EA charging and the integration of Photo-Voltaic (PV) and BESS into an MG. Furthermore, the integration of hydrogen in airport operations is discussed in, for example, [9]. Hence, a research gap remains in the development of EMS solutions for airport TE-MGs that coordinate the operation of electrical and thermal resources while considering e-plane charging demand and multi-zone building thermal dynamics.

The energy management of TE-MGs has been extensively studied across diverse applications using various optimization techniques. Among these, MPC is widely recognized as an effective control strategy due to its robustness and sample efficiency. Thus, it has been employed to account for uncertainties in EMS problems (e.g., [4], [10], [11]), and has also been used as benchmark to evaluate the performance of novel proposed stochastic techniques [12]–[14]. However, MPC requires model-based representations of the system components, which can be difficult to ascertain, and may involve high computational costs. For example, when thermal comfort is considered, determining the building's thermal parameters, i.e., thermal resistances and capacitances, is a challenging task, especially in multi-zone thermal models. Hence, MPC-based EMS optimization often relies on simplified models. However, even when such models adequately capture the system dynamics, they may still fail to adapt to changing conditions due to system modifications or component aging [12].

To overcome the difficulties with optimization-based EMS, Reinforcement Learning (RL) has been widely investigated in recent years as an alternative for solving EMS problems, as it does not require an explicit model of the system, with an agent learning through direct interaction with the environment and generating its own training data via experience. This makes RL especially valuable for situations where data is limited, expensive, and/or constantly changing [15]. At its core, RL focuses on learning how to map situations to actions in order to maximize long-term rewards by discovering effective strategies, i.e., policies, by trial and error [16]. In traditional RL, key functions such as the policy and the value function are often represented using lookup tables. However, in many real-world applications, the complexity and dimensionality of the state and action spaces make such explicit representations infeasible [15]. Thus, Deep Reinforcement Learning (DRL) has emerged as an extension of classical RL to enable the application of learning-based approaches to complex, high-dimensional problems, such as the energy management of TE-MGs. For instance, scheduling strategies for appliances and EVs in residential systems have been studied in [17]–[20]. Researchers in, for example, [21] and [22], have explored the optimal operation of multi-MG systems, where multi-agent approaches are typically used, with each agent optimizing the operation of an individual MG. Another example of multi-agent DRL is presented in [23], where agents independently control a BESS and fan coil units for room temperature control in a building MG. Coordinating multiple agents, however, adds layers of complexity and poses scalability concerns as the system grows.

Although DRL offers advantages in managing computational complexity, several challenges remain with respect to the effective handling of continuous and hybrid action spaces that involve both discrete and continuous variables, accurately capturing system dynamics, designing meaningful reward functions, and adequately tuning hyperparameters. Thus, to address these issues, [13] proposes a multi-stage reward mechanism and expert intervention, which aims to prevent the agent from being trapped in sub-optimal strategies. Researchers in [21] and [24] have considered including physics-informed reward shaping to speed up the agent's learning process, while [25] introduces a diffusion model to obtain more accurate data features, and a two-step reward function. Furthermore, hybrid methods that combine RL with traditional optimization approaches, such as MPC, or Mixed Integer Quadratic Programming (MIQP), have been developed to leverage the complementary strengths of data-driven learning and model-based control [14], [26]. A well-known limitation of conventional RL approaches is their inability to explicitly handle system constraints. Hence, MPC has been used as a supporting mechanism to enforce constraints [14], [27]. Beyond this hybrid approach, various methods have been proposed to address constraint handling directly within RL. For instance, the application of strict physical constraints [23], and primal-dual optimization in constrained Markov Decision Process (MDP) formulations [26], [28], have been introduced. However, these approaches face several challenges, including convergence instability, sensitivity to reward–constraint trade-offs, and a lack of strong safety guarantees. These limitations have motivated the exploration of alternative safe RL strategies that enforce stricter guarantees. Thus, in [29], authors propose incorporating hard constraints that are formulated independently of the underlying RL control technique, enabling robust safety enforcement regardless of the learning algorithm, while in [30], safety constraints are embedded directly into the RL environment through a constrained MDP formulation, which is solved using Interior-Point Policy Optimization (IPO) to ensure safe and reliable policy learning.

With the exception of [24], [27], [30], and [31], most of the proposed RL-based techniques do not incorporate building thermal models to capture the underlying thermal dynamics, and instead, forecasted thermal demand profiles are typically used as inputs to the EMS. While this simplification reduces modeling complexity, it decouples thermal behavior from control actions, limiting the system's ability to account for thermal inertia and temperature constraints. Furthermore, among the works that do consider building thermal modeling, only [31] addresses a multi-zone building configuration; however, the proposed EMS is not applied within a MG framework, and focuses solely on room temperature control.

Based on the preceding literature review, the following four main gaps can be identified:

- Existing airport MG studies have mainly focused on planning, scheduling problems, hydrogen integration, or the provision of ancillary services, while EMS formulations for airport TE-MGs with coupled electrical and multi-zone thermal dynamics are limited.
- Many DRL-based EMS studies use thermal demand profiles rather than temperature-based building thermal models, which limits their ability to capture thermal inertia, inter-zone heat exchange, and comfort constraints.
- Constraint handling in DRL-based EMSs is challenging, particularly for maintaining indoor temperature within prescribed comfort limits in multi-zone buildings.
- Comparative evaluations of DRL-based EMSs against optimization-based benchmarks are often limited to cost- or reward-based performance indicators, with less emphasis on system-level operational metrics such as BESS behavior and indoor temperature regulation.

Motivated by these limitations, this paper makes the following contributions:

- Develop a novel DRL-based EMS framework for the optimal operation of an airport TE-MG incorporating real-world data for load demand, solar radiation profiles, ambient temperature, and e-plane charging, employing a control formulation that combines continuous control variables and discrete operational decisions to realistically capture the operational characteristics of the TE-MG.
- Integrate physics-based dynamic constraints within the DRL environment, implemented through state-dependent action clipping, to ensure safe and physically consistent control of both the BESS and indoor temperature, and implement a multi-zone building thermal model to properly represent heat dynamics and inter-zone heat exchange, using an implicit method to efficiently handle variable coupling.
- Conduct a systematic comparative evaluation against an MPC-based benchmark to showcase the performance of the DRL agent in an actual airport MG, assessing not only cost performance but also system-level operational metrics, including BESS behavior, multi-zone indoor temperature regulation, and real-time computation costs.

The rest of the paper is organized as follows: Section II provides a background overview of MG EMS and related thermal systems modeling, and outlines the formulation of DRL as an MDP. Section III describes the proposed DRL framework for the energy management of TE-MGs, including the definition of the action and observation spaces, environment setup, and reward design. Section IV comprises the application of the proposed DRL framework to an actual hangar building currently under deployment at the Waterloo Wellington Flight Centre (WWFC). Furthermore, the proposed framework is benchmarked against an MPC-based EMS, using appropriate metrics to evaluate the DRL agent with respect to BESS management and multi-zone indoor temperature control requirements, while the training performance is compared against a reward-based DRL method. Finally, Section V provides the summary and conclusions of the presented studies.

II. Background

A. EMS for Thermo-Electrical Microgrids

The EMS constitutes the supervisory optimization layer in MGs, applicable to both grid-connected and islanded operation, and may be implemented through centralized or decentralized architectures [2]. This paper focuses on implementing a centralized EMS scheme, wherein a central controller determines the dispatch of available resources based on forecast information and predefined objectives. When implemented with DRL, the agent functions as this central controller, observing the state of the MG, selecting coordinated control actions for all resources, and continuously improving its policy through reward-driven interactions with the environment [22].

The modeling of Thermal Energy Systems (TESs) in EMS applications can be effectively achieved using the Thermal Equivalent Circuit (TEC) method, which establishes an analogy between thermal and electrical variables, enabling an accurate representation of thermal dynamics through an equivalent lumped-parameter electrical circuit model [32], wherein thermal power and temperature correspond to current and voltage, respectively. The resulting equivalent circuit can subsequently be solved to derive the differential equations describing relevant thermodynamic processes. According to [10], the indoor air temperature dynamics can be effectively captured using a first-order model that accounts for the main heat gains in each room i. However, as additional elements are considered, such as heat exchange with walls or multi-zone configurations, the order of the model naturally expands. Hence, as per [4], the EMS constraints associated with the

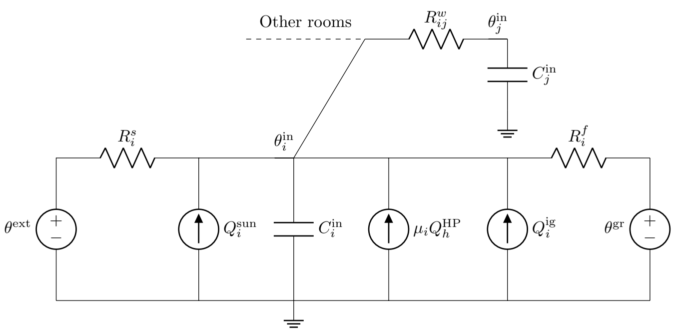

Fig. 1. Thermal equivalent circuit of room i [4].

thermal dynamics of a given room i can be derived from the building thermal model illustrated in Fig. 1, as follows:

$$C_i^{\text{in}} \frac{d\theta_i^{\text{in}}}{dt} = \mu_i Q_h^{\text{HP}} - Q_{ij}^{\text{trf}} - Q_i^{\text{env}} - Q_i^{\text{gr}} + Q_i^{\text{sun}} + Q_i^{\text{ig}} \quad (1)$$

where all parameters and variables are defined in the Nomenclature section for this and other equations. This equation defines the indoor temperature dynamics in room i, resulting from the sum of various thermal power contributions and losses, and is represented in discrete form within the EMS model, employing a time step on the order of minutes. Although the thermal capacitance, and thus the storage capability, of walls, floors, and ceiling is not explicitly modeled, their thermal conductivity values are properly accounted for. Thus, due to the high thermal conductivity of partition walls, their energy storage capacity is low and typically neglected. Conversely, the thermal capacitance of insulated surfaces with low conductivity, and thus higher thermal storage potential, is typically aggregated into the thermal capacitance of the rooms to reduce the number of model variables.

B. RL Framework Overview

The energy management problem can be formulated within a DRL framework as an MDP. Thus, the Markov property asserts that the probability of transitioning to the next state depends solely on the current state and action, and not on any prior states [16]. Formally, an MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, representing the environment's status at each time step; $\mathcal{A}$ is the action space, denoting the available control actions; $\mathcal{P}(s_{k+1} \mid s_k, a_k)$ is the state transition probability function, which describes the system's evolution from state s_k to s_{k+1} under action a_k; $\mathcal{R}(s_k, a_k, s_{k+1})$ is the reward function, which quantifies the immediate cost or benefit resulting from taking action a_k in state s_k and transitioning to the next state s_{k+1}; and $\gamma \in [0, 1]$ is the discount factor, which balances the importance of future versus immediate rewards. As per [16], at each time step, the agent follows a policy π that defines a mapping from states to probabilities of selecting each possible action.

RL algorithms can be broadly categorized into the following classes: value-based, policy-based, and actor-critic methods. Value-based methods, such as Q-learning, estimate the expected return of each action and derive a policy implicitly; policy-based methods directly optimize the policy without relying on value estimates; and actor-critic methods combine the benefits of both approaches. Thus, the actor updates the policy, while the critic evaluates the value function to guide the learning process. The current paper adopts the Proximal Policy Optimization (PPO) algorithm, which is a state-of-the-art actor-critic method known for its robustness and stable performance in continuous control problems [33]. PPO optimizes a clipped surrogate objective function L, which limits the extent of policy updates to prevent instability during training as follows:

$$L^{\text{CLIP}}(\phi) = \mathbb{E}_k \left[\min \left(p_k(\phi) \hat{A}_k,\ \tilde{p}_k \hat{A}_k \right) \right] \quad (2)$$

where ϕ denotes the parameters of the policy network; $\mathbb{E}_k[\cdot]$ represents the empirical expectation over a batch of sampled time steps k; $p_k(\phi) = \frac{\pi_\phi(a_k|s_k)}{\pi_{\phi_{\text{old}}}(a_k|s_k)}$ is the probability ratio between the new and old policies; and $\hat{A}_k$ is the advantage estimate, which quantifies how much better or worse an action is compared to the expected value of the state, guiding the policy to favor actions that lead to higher-than-expected returns and thus enhancing the agent's learning efficiency. The term $\tilde{p}_k$ is the clipped probability ratio constrained to the interval $[1-\epsilon, 1+\epsilon]$, where ϵ controls the extent to which the new policy is allowed to deviate from the old one.

To address constraint satisfaction in RL frameworks, various strategies have been proposed, including reward shaping, constrained-MDPs solved via Lagrangian multipliers, and enforcing feasibility directly in the environment through action clipping. In this work, the latter approach is adopted, implementing physics-based action clipping to ensure that the actions selected by the agent are projected onto feasible operating ranges. This method is computationally efficient and can improve training stability by reducing infeasible exploration and restricting the actions applied within the environment to feasible operating ranges, making it particularly well suited for continuous control problems such as energy management in MGs.

III. Proposed RL-Based EMS Framework

This section describes the DRL framework proposed for the EMS of a multi-zone TE-MG for an airport building, which incorporates a detailed model to accurately capture room thermal dynamics. Electric power demand is assumed to be supplied from the main grid, PV generation units, and the BESS, whereas thermal energy is provided by Heat Pumps (HPs) and an Electric Heater (EH) for the aircraft hangar.

A. Action Space

The proposed DRL framework models the interaction of a single agent with the environment to control the operation of the BESS, multiple HPs, and an EH at each time step k. Thus, the multidimensional action vector comprising the various control components can be defined as follows:

$$a_k = \left[a_k^{\text{bat}},\ a_{h,k}^{\text{HP}},\ a_k^{\text{EH}} \right] \quad (3)$$

where a_k^{bat} and $a_{h,k}^{\text{HP}}$ can take positive or negative values representing charging and discharging for the former, and heating and cooling for the latter, respectively; and a_k^{EH} is the discrete control signal for the EH, taking values $\{0,1\}$ to indicate off and on states, respectively. In the implemented PPO framework, the policy outputs a continuous multidimensional action vector. Thus, although the physical EH command is binary, it is represented at the policy-output level by an action component bounded between 0 and 1 based on a continuous relaxation approach, which is then rounded in the environment to obtain the required ON/OFF command. Therefore, the hybrid nature of the control problem is handled through continuous policy outputs and environment-level action processing, rather than through a separate discrete-action policy.

B. State Space

The state space of the TE-MG environment must account for all the system variables the agent needs to observe in order to achieve an optimal operation. In this context, the state of the system components at each time step is defined as follows:

$$s_k = \left[k, P_k^{\mathrm{PV}}, P_k^{\mathrm{d}}, P_k^{\mathrm{ep}}, c_k^{\mathrm{gp}}, SOC_k, \theta_k^{\mathrm{ext}}, \theta_{i,k}^{\mathrm{in}}, \underline{\theta}_{i,k}^{\mathrm{in}}\right] \quad (4)$$

which encompasses input data, namely, PV generation, building and e-plane charging power demand, Time of Use (TOU), and ambient temperature, as well as a time-varying temperature limit $\underline{\theta}_{i,k}^{\mathrm{in}}$ based on room comfort considerations. It also incorporates system variables, i.e., BESS State of Charge (SOC) and room temperatures. Note that only the lower temperature limit is included, since it is assumed to vary during the day, whereas the upper limit is considered to be fixed, as is typically the case. In the implemented observation space, some of the exogenous variables in (4) are provided to the agent, namely, PV generation, demand, ambient temperature, TOU electricity price, and time-varying temperature limits, using a rolling look-ahead window whose length can be adjusted according to the desired forecast horizon. Thus, at each decision step, the agent observes the current value and the next forecasted values within the defined look-ahead window.

C. RL Environment

The RL environment comprises the modeling of the EMS elements, ensuring that the system variables are properly updated whenever an agent takes actions based on the observations at each time step, since measurements were not available for the TE-MG being considered. In practical applications, system observations can be obtained directly, without relying on equation-based models, thus reducing modeling complexity and errors. However, as highlighted throughout this section, certain equations remain necessary to enforce operational constraints, such as the power and capacity limits governing BESS and HP operation. It is important to highlight that, although the TE-MG model was used to develop and demonstrate the proposed DRL-based EMS approach, it also allowed comparison and validation of the latter with respect to an optimization-based MPC EMS.

The proposed model assumes that the electric demand, associated with the building, e-plane, HPs, and EH, is fully satisfied by the available generation. Hence, the power balance is defined as follows:

$$P_k^{\mathrm{gp}} = P_k^{\mathrm{d}} + P_k^{\mathrm{ep}} + P_k^{\mathrm{bat}} - P_k^{\mathrm{PV}} + \sum_{h\in H} P_{h,k}^{\mathrm{HP}} + a_k^{\mathrm{EH}} P^{\mathrm{EH}} \quad (5)$$

where P^{EH} represents the rated power of the EH, and P_k^{gp} corresponds to the power exchanged with the grid. In a real-world application, $P_k^{\mathrm{d}}, P_k^{\mathrm{ep}}$, and P_k^{PV} would be measured directly from load consumption and PV generation; $P_k^{\mathrm{bat}}, P_{h,k}^{\mathrm{HP}}$, and a_k^{EH} would be determined by the DRL agent; and the resulting P_k^{gp} would be measured at the Point of Common Coupling (PCC).

The BESS operation at each time step k is modeled as follows:

$$P_k^{\mathrm{ch}} = \max\left(0, P_k^{\mathrm{bat}}\right), \quad P_k^{\mathrm{dch}} = \max\left(0, -P_k^{\mathrm{bat}}\right) \quad (6)$$

$$SOC_{k+1} - SOC_k = \left(P_k^{\mathrm{ch}}\eta^{\mathrm{ch}} - \frac{P_k^{\mathrm{dch}}}{\eta^{\mathrm{dch}}}\right)\Delta t_k \quad (7)$$

$$\underline{SOC} \leq SOC_k \leq \overline{SOC} \quad (8)$$

$$\Phi_k(\delta_k) = 5.24 \times 10^{-4}\delta_k^{2.03} \quad (9)$$

where (6) decomposes the BESS action a_k^{bat} into two non-negative components to separately model charging and discharging power, i.e., P_k^{ch} and P_k^{dch}, respectively. Equation (7) represents the BESS energy balance; and the SOC limits are modeled in (8). Based on [34], $\Phi_k(\delta_k)$ in (9) represents BESS degradation, as a function of $\delta_k = 1 - SOC_k$, which is a variable indicating the BESS Depth of Discharge (DOD).

Finally, the MG thermal dynamics, which ensure both comfort compliance and reliable operation, can be represented as follows:

$$\begin{aligned} C_i^{\mathrm{in}}\frac{\theta_{i,k}^{\mathrm{in}} - \theta_{i,k-1}^{\mathrm{in}}}{3600\Delta t_k} &= \mu_i Q_{h,k}^{\mathrm{HP}} + Q_{i,k}^{\mathrm{sun}} - Q_{i,k}^{\mathrm{env}} - Q_{i,k}^{\mathrm{vent}} \\ &- \sum_{j\in\mathcal{I}} Q_{ij,k}^{\mathrm{trf}} + Q_{i,k}^{\mathrm{ig}} - Q_{i,k}^{\mathrm{gr}} \quad \forall i \in \mathcal{I}\backslash\{ph\},\ \forall k \in T \end{aligned} \quad (10)$$

$$\begin{aligned} C_i^{\mathrm{in}}\frac{\theta_{i,k}^{\mathrm{in}} - \theta_{i,k-1}^{\mathrm{in}}}{3600\Delta t_k} &= a_k^{\mathrm{EH}} P^{\mathrm{EH}} + Q_{i,k}^{\mathrm{sun}} - Q_{i,k}^{\mathrm{env}} - Q_{i,k}^{\mathrm{vent}} \\ &- \sum_{j\in\mathcal{I}} Q_{ij,k}^{\mathrm{trf}} + Q_{i,k}^{\mathrm{ig}} - Q_{i,k}^{\mathrm{gr}} \quad \forall i \in \{ph\},\ \forall k \in T \end{aligned} \quad (11)$$

$$Q_{ij,k}^{\mathrm{trf}} = \sum_{\mathrm{w}\in W} U_{ij}^{\mathrm{w}} A_{ij}^{\mathrm{w}}\left(\theta_{i,k}^{\mathrm{in}} - \theta_{j,k}^{\mathrm{in}}\right) \ \forall i,j \in \mathcal{I},\ \forall k \in T \quad (12)$$

$$Q_{i,k}^{\mathrm{gr}} = U_i^{\mathrm{f}} A_i^f\left(\theta_{i,k}^{\mathrm{in}} - \theta^{\mathrm{gr}}\right) \quad \forall i \in \mathcal{I}\backslash\{ng\},\ \forall k \in T \quad (13)$$

$$Q_{h,k}^{\mathrm{HP}} = \psi_h^{\mathrm{HP}}\frac{\theta_h^{\mathrm{HP,out}}}{\theta_h^{\mathrm{HP,out}} - \theta_k^{\mathrm{ext}}} P_{h,k}^{\mathrm{HP}} \quad \forall h \in H,\ \forall k \in T \quad (14)$$

$$Q_{i,k}^{\mathrm{sun}} = \sum_{e\in\mathcal{B}} \sigma_{e,i} I_k A_{e,i}\xi_{e,i}\tau_{e,i} \quad \forall i \in \mathcal{I},\ \forall k \in T \quad (15)$$

$$Q_{i,k}^{\mathrm{env}} = \sum_{e\in\mathcal{B}} U_{e,i} A_{e,i}\left(\theta_{i,k}^{\mathrm{in}} - \theta_k^{\mathrm{ext}}\right) \quad \forall i \in \mathcal{I},\ \forall k \in T \quad (16)$$

$$Q_{i,k}^{\mathrm{vent}} = \frac{v_{i,k}^{\mathrm{rate}}\rho^{\mathrm{air}} Vol_i\, c^{\mathrm{air}}\left(\theta_{i,k}^{\mathrm{in}} - \theta_k^{\mathrm{ext}}\right)}{3600} \quad \forall i \in \mathcal{I},\ \forall k \in T \quad (17)$$

$$Q^{\mathrm{ig}}_{i,k} = \mathrm{Np}_{i,k} Q^{\mathrm{oc}}_{i} \mathrm{CLF}_{k} + Q^{\mathrm{lg}}_{i,k} + Q^{\mathrm{eq}}_{i,k} \quad \forall i \in \mathcal{I},\ \forall k \in T \tag{18}$$

where the indoor air thermal dynamics are modeled in (10) and (11), considering the thermal contribution from HPs and the EH, respectively, and that most of EH electric power is converted to heat. The heat exchange occurring between adjacent rooms and between the rooms and the ground is represented by (12) and (13), respectively, with the ground temperature assumed constant over the course of the day owing to the soil's high thermal inertia and the time horizon of the EMS. The thermal power provided by the HPs is shown in (14), with its corresponding Coefficient of Performance (COP) at each time step $COP_k = \psi^{\mathrm{HP}}_h \frac{\theta^{\mathrm{HP,out}}_h}{\theta^{\mathrm{HP,out}}_h - \theta^{\mathrm{ext}}_k}$, which depends on the temperature of the HP outlet, the outside temperature, and the coefficient reduction ψ that represents the deviation from ideal efficiency. Equations (15) to (17) describe the thermal contributions from solar irradiation, building-envelope heat transfer, and ventilation, respectively, while (18) accounts for internal heat gains. In an actual deployment, equations (7), (12), (13), (16), and (17) can be replaced by direct sensor readings (e.g., wall heat flux sensors, ventilation flow meters), whereas (14), (15), and (18) may be estimated using partial measurements combined with model-based representations.

D. Physics-Based Dynamic Constraints

To ensure constraint satisfaction, the agent's raw actions are restricted in the DRL environment using physics-based dynamic constraints for the thermal system and BESS SOC derived from the system's physical characteristics and limits. These constraints are implemented through state-dependent action clipping, where the feasible operating bounds are updated at each time step based on the current system conditions. In particular, the feasible HP bounds are determined from the multi-zone thermal model and indoor temperature limits, while the feasible BESS charging/discharging bounds are determined from the SOC dynamics, efficiency losses, and SOC limits, guiding the agent toward physically valid and deployable control policies. The corresponding bounds and clipping operations are defined as follows:

$$\begin{aligned}\underline{Q}^{\mathrm{HP}}_{h,k} = C^{\mathrm{in}}_i \frac{\underline{\theta}^{\mathrm{in}}_{i,k} - \theta^{\mathrm{in}}_{i,k-1}}{3600\Delta t_k} - Q^{\mathrm{sun}}_{i,k} + Q^{\mathrm{env}}_{i,k} + Q^{\mathrm{vent}}_{i,k} \\ + \sum_{j\in\mathcal{I}} Q^{\mathrm{trf}}_{ij,k} - Q^{\mathrm{ig}}_{i,k} + Q^{\mathrm{gr}}_{i,k} \quad \forall i \in \mathcal{I}\backslash\{nc\}\end{aligned} \tag{19}$$

$$\begin{aligned}\overline{Q}^{\mathrm{HP}}_{h,k} = C^{\mathrm{in}}_i \frac{\overline{\theta}^{\mathrm{in}}_{i,k} - \theta^{\mathrm{in}}_{i,k-1}}{3600\Delta t_k} - Q^{\mathrm{sun}}_{i,k} + Q^{\mathrm{env}}_{i,k} + Q^{\mathrm{vent}}_{i,k} \\ + \sum_{j\in\mathcal{I}} Q^{\mathrm{trf}}_{ij,k} - Q^{\mathrm{ig}}_{i,k} + Q^{\mathrm{gr}}_{i,k} \quad \forall i \in \mathcal{I}\backslash\{nc\}\end{aligned} \tag{20}$$

$$\begin{aligned}P^{\mathrm{HP}}_{h,k} = \mathrm{clip}\Bigg(a^{\mathrm{HP}}_{h,k},\ \max\left(\underline{P}^{\mathrm{HP}}_h,\ \frac{\underline{Q}^{\mathrm{HP}}_{h,k}}{COP_k}\right), \\ \min\left(\overline{P}^{\mathrm{HP}}_h,\ \frac{\overline{Q}^{\mathrm{HP}}_{h,k}}{COP_k}\right)\Bigg)\end{aligned} \tag{21}$$

where the clipping operator is defined as $\mathrm{clip}(a, \underline{a}, \overline{a}) = \min(\max(a, \underline{a}), \overline{a})$, which projects a onto the interval $[\underline{a}, \overline{a}]$, and (19) and (20) define the minimum and maximum HP thermal power limits that ensure indoor thermal comfort. Notably, the HP limits are divided by the COP to define the physics-based dynamic upper and lower bounds for HP operation in (21), which allows for operational flexibility, as the maximum electrical power supplied by the HP is constrained by the HP's rated power limits. Similarly, the dynamic limits for BESS operation can be determined as follows:

$$\overline{P}^{\mathrm{bat}}_k = \max\left(0,\ \min\left(\overline{P}^{\mathrm{bat}},\ \frac{\overline{SOC} - SOC_k}{\eta^{\mathrm{ch}}\,\Delta t_k}\right)\right) \tag{22}$$

$$\underline{P}^{\mathrm{bat}}_k = \min\left(0,\ \max\left(\underline{P}^{\mathrm{bat}},\ \frac{(\underline{SOC} - SOC_k)\,\eta^{\mathrm{dch}}}{\Delta t_k}\right)\right) \tag{23}$$

$$P^{\mathrm{bat}}_k = \mathrm{clip}\left(a^{\mathrm{bat}}_k,\ \underline{P}^{\mathrm{bat}}_k,\ \overline{P}^{\mathrm{bat}}_k\right) \tag{24}$$

where the second terms in (22) and (23) impose physics-based dynamic constraints to prevent overcharging or over-discharging, accounting for the measured SOC_k and considered efficiency losses. Finally, (24) ensures that the selected action a^{bat}_k is clipped to the feasible range at each time step.

It should be noted that feasibility is enforced at the environment level by processing the raw actions before the system states are updated. Therefore, the learned policy itself is not required to independently generate feasible actions at all times; rather, physically valid operation is maintained through the embedded action-processing mechanism.

E. Multi-Zone Thermal Modeling

The thermal modeling of the building is used here to obtain the required room temperatures, which in an actual deployment would be directly measured and thus not included in a model-based DRL environment. In this context, the thermal dynamics of a building with m rooms can be modeled using a multi-zone network model, where at each time step k, the room temperatures $\boldsymbol{\theta}_k \in \mathbb{R}^m$ can be obtained by solving the following system of equations:

$$\mathbf{M} \cdot \boldsymbol{\theta}_k = \mathbf{b}_k \tag{25}$$

This is solved here using the implicit Euler method, which ensures numerical stability despite the interdependencies between zones, thus resulting in the following M and b terms:

$$M_{ii} = 1 + \frac{3600\Delta t_k}{C^{\mathrm{in}}_i} \sum_{j\in\mathcal{I}\backslash\{i\}} U^{\mathrm{w}}_{ij} A^{\mathrm{w}}_{ij} \tag{26}$$

$$M_{ij} = -\frac{3600\Delta t_k}{C^{\mathrm{in}}_i} U^{\mathrm{w}}_{ij} A^{\mathrm{w}}_{ij} \quad \forall j \in \mathcal{I},\ i \neq j \tag{27}$$

$$\begin{aligned}b_{i,k} = \theta_{i,k-1} + \frac{3600\Delta t_k}{C^{\mathrm{in}}_i}\Big(\mu_i Q^{\mathrm{HP}}_{h,k} + Q^{\mathrm{EH}}_{i,k} + Q^{\mathrm{sun}}_{i,k} - Q^{\mathrm{env}}_{i,k} \\ - Q^{\mathrm{vent}}_{i,k} + Q^{\mathrm{ig}}_{i,k} - Q^{\mathrm{gr}}_{i,k}\Big)\end{aligned} \tag{28}$$

F. Reward Function

The objective of the proposed DRL-based EMS is to learn a policy that reduces grid operating cost while maintaining indoor thermal comfort and physically meaningful BESS operation over the daily operating horizon. In contrast to an MPC-based approach, where the EMS objective is optimized explicitly at each control interval, the DRL-based EMS represents this objective through the reward function and optimizes it during training by maximizing the following expected discounted cumulative reward:

$$J(\pi_\phi) = \mathbb{E}_{\pi_\phi}\left[\sum_{k=0}^{T-1} \gamma^k r_k\right] \tag{29}$$

where π_ϕ denotes the policy parameterized by ϕ, and r_k is the reward obtained at time step k, which is designed here to penalize grid energy exchange costs, BESS degradation, indoor temperature violations, and terminal SOC deviations. Thus, the following penalty-based structure discourages suboptimal operating behavior and guides the agent toward economically efficient and physically consistent operation:

$$r_k = -c_k^{\text{gp}} P_k^{\text{gp}} \Delta t_k - \lambda_1 c_k^{\text{BESS}} - \lambda_2 c_k^{\text{temp}} - c_k^{\text{SOC}} \tag{30}$$

$$c_k^{\text{BESS}} = E_n RC\Phi_k \tag{31}$$

$$c_k^{\text{temp}} = \max\left(0, \underline{\theta}_{\{ph\}}^{\text{in}} - \theta_{\{ph\},k}^{\text{in}}\right)^2 \tag{32}$$

$$c_k^{\text{SOC}} = \begin{cases} \left|SOC_k - SOC^{\text{tgt}}\right|, & \text{if } k = T-1 \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

In the reward function (30), the first term represents the grid energy exchange cost, where c_k^{gp} is the grid electricity price, P_k^{gp} is the grid power determined by the electrical power balance, and Δt_k is the time-step duration. The coefficients λ_1 and λ_2 balance the relative importance of battery degradation and indoor temperature violations, respectively. These coefficients should be selected strategically to maintain an appropriate trade-off among the different reward components, since excessively large penalty values may dominate the reward signal and lead the agent toward overly conservative or undesirable policies. For instance, a large temperature-violation penalty may cause the agent to prioritize thermal comfort at the expense of economically meaningful grid and BESS operation, while a large BESS degradation penalty may discourage useful battery cycling and limit arbitrage opportunities. Conversely, small penalty values may lead to poor BESS management or comfort violations. Equation (31) defines the cost associated with degradation of a lithium-ion BESS c_k^{BESS}, as per [4]. The penalty term for temperature violations is modeled in (32) and is applied exclusively to the planes' hangar, as the temperature in other rooms remains within limits due to the imposed physics-based constraints. Finally, c_k^{SOC} allows the BESS to reach a target SOC^{tgt} at the end of the considered horizon.

Note that the variables appearing in the reward function (30) are directly linked to the mathematical formulations presented in the previous subsections. Specifically, P_k^{gp} is obtained from the electrical power balance in (5); the battery degradation cost is based on the degradation function defined in (9); SOC_k evolves according to the BESS dynamics defined in (7); and $\theta_{\{ph\},k}^{\text{in}}$ is obtained from the solution of the multi-zone thermal model represented in (25). Thus, the mathematical formulation defines the state transitions and feasible operating conditions of the TE-MG, while the reward function evaluates the resulting control actions. The overall training process of the agent under the proposed DRL framework is summarized in Algorithm 1.

Algorithm 1 RL-Based EMS Training Framework

Initialize policy π_ϕ and EMS environment
for each iteration **do**
  for $k = 0$ to $T-1$ **do**
    Observe s_k
    Sample action $a_k \sim \pi_\phi(a \mid s_k)$
    Apply physics-based constraints: **Clip** a_k
    Control BESS, HPs, EH with a_k
    Compute $P_k^{\text{gp}},\ SOC_k$
    Solve multi-zone thermal model for $\theta_{i,k}^{\text{in}}$
    Compute reward r_k
    Observe next state s_{k+1}
    Store transition (s_k, a_k, r_k, s_{k+1})
  end for
  Update policy π_ϕ with collected data
end for

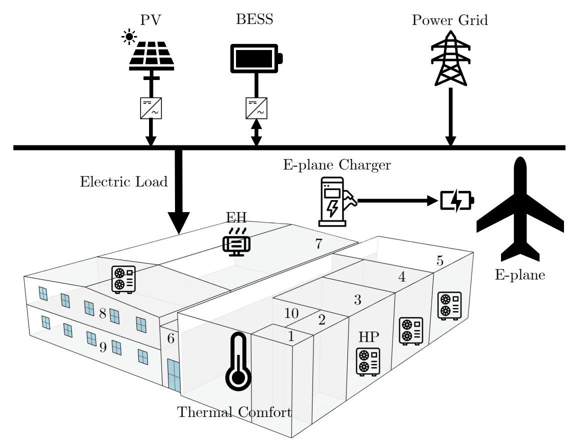

Fig. 2. Microgrid at WWFC Hangar 7.

IV. Results and Discussion

The proposed DRL framework is applied to the WWFC airport MG being developed at Waterloo's airport in Ontario, which is illustrated in Fig. 2. This system consists of a PV plant, which is the primary RES in airport MGs due to restrictions on wind generation; a BESS; four HPs for indoor temperature control in building rooms, which include 3 flight simulator rooms as well as office and classroom spaces; an EH used to heat the planes' hangar; and an e-plane charger. The building at WWFC is divided into ten distinct zones, as portrayed in Table I, with their corresponding thermal characteristic provided in Table II, and associated HP thermal distribution factors in Table III, based on [4]. This allows to simulate the building thermal dynamics, since actual measurements for temperatures and thermal contributions are not yet available. Internal heat gains are represented by considering the expected room occupancy, as per [35], along with the sensible heat from the electrical equipment and lighting planned for installation in the building.

TABLE I
WWFC HANGAR 7 ROOM INDEX AND DIMENSIONS [4]

Room i	Description	Area [m^2]	Volume [m^3]
1	Electric room	6.5	37.3
2	Restrooms	8.5	48.5
3	Flight simulator 1	37.2	213.3
4	Flight simulator 2	37.2	213.3
5	Flight simulator 3	69.7	400.0
6	Corridor 1	52.0	222.0
7	Planes' hangar	278.7	1444.2
8	Classrooms	111.5	288.8
9	Offices	111.5	288.8
10	Corridor 2	77.9	447.4

TABLE II
VALUES FOR THERMAL PARAMETERS AT WWFC HANGAR 7 [4]

	Building [m^{2}K/W]	Hangar [m^{2}K/W]	Corridor [m^{2}K/W]
Outside walls	4.93	3.17	4.93
Roof	4.93	3.52	4.93
Partition walls	0.25	3.17	4.93
Ground slab	2.38	2.38	2.38

TABLE III
HP THERMAL DISTRIBUTION FACTORS

	μ_1	μ_2	μ_6	μ_8	μ_9	μ_{10}
Summer	0	0	0.1	0.3	0.3	0.3
Winter	0.07	0.08	0.1	0.3	0.15	0.3

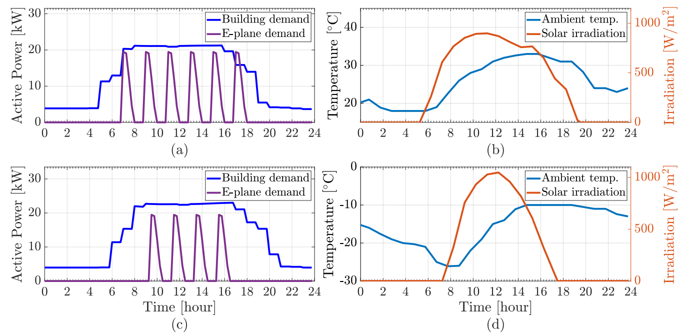

Fig. 3. Electric demand and weather conditions for WWFC Hangar 7 for the hottest summer day (a) and (b) and the coldest winter day (c) and (d).

The EMS in airport MGs must adequately account for the e-plane power demand associated with the required charging operations. Hence, this work incorporates real-world e-plane power demand data obtained from the flight activity of the Pipistrel Velis Electro at the WWFC. Based on daylight availability and airport traffic, the e-plane is assumed to undergo six and five charging sessions per day during summer and winter, respectively. However, for the coldest winter day, only four charging operations are considered, as the e-plane can only fly when ambient temperatures exceed -20°C [36]. Since the e-plane demand profile assumes a minimum SOC of 30% at touchdown, approximately 14 kWh of energy is required per charging session to recharge the 24.8 kWh e-plane.

The DRL-based EMS considers a daily time horizon of $T = 24$ h, discretized into uniform time steps of duration $\Delta t = 15$ min, similar to the optimization-based EMS proposed in [4], which is used here for comparison and validation purposes. The DRL agent aims to learn a control strategy that optimizes performance over this horizon, balancing immediate objectives, such as cost and thermal comfort at each step, with longer-term constraints associated with BESS operation.

Due to the substantially different seasonal operating conditions of the airport TE-MG, separate DRL policies are trained for warm and cold days. This training strategy avoids forcing a single policy to approximate two significantly different operating regimes, since cold and warm operation differ in terms of ambient temperature, PV generation, thermal-control requirements, and electricity demand. During training, the representative daily profiles are repeatedly sampled, and stochastic perturbations are applied to the forecasted exogenous inputs over a 1-hour look-ahead window to expose the agent to variations around the selected daily operating patterns. The trained DRL policies are also evaluated under imperfect forecast information by applying normally distributed perturbations to the 1-hour look-ahead observations during testing.

The trained DRL policies are configuration-specific and are not intended to automatically adapt to structural changes in the airport hangar TE-MG, given the fact that the physical behavior of the thermal system and battery SOC are embedded in the DRL environment for constraint enforcement. Changes in component ratings, installed DERs, thermal-zone layout, or e-plane charging infrastructure may require updating the state and action definitions, thermal system component and BESS SOC models, temperature and SOC operational constraints, and reward terms. In such cases, the modular DRL environment can be updated accordingly, and the policy can be retrained or fine-tuned offline using the revised physical model of the system and operational data. Thus, the proposed approach should be interpreted as a reusable DRL framework with physics-based constraint enforcement through the environment and action-processing mechanism, rather than as a fixed policy that is universally transferable to all possible future configurations.

The detailed time-series results presented in Section IV-A focus on two boundary operating scenarios, corresponding to the hottest summer day and the coldest winter day based on hourly averaged meteorological data for Ontario, Canada, from 2007 to 2021 [4]. These two scenarios are selected to evaluate the controller under boundary seasonal conditions with high cooling or heating requirements. The e-plane and building power demand, together with the environmental conditions for these boundary days, are depicted in Fig. 3. Furthermore, the lower temperature setpoint is defined as 21°C for working hours, which extend from 8 am to 9 pm, and 15°C during the remaining hours, while a constant setpoint of 10°C is assumed for the aircraft hangar for the whole day. The maximum temperature is set to 24°C for all rooms except the hangar, for which no cooling system is available.

The hyperparameters used for training were selected using Optuna [37], followed by a manual fine-tuning process. Hence, the architecture of the proposed DRL framework is shown in Fig. 4, with a discount factor of $\gamma = 0.996$, an entropy coefficient of 0.01 to encourage exploration, a learning rate of 3×10^{-5}, and 15 training epochs. Training was conducted on a workstation equipped with two Intel® Xeon® Silver 4214

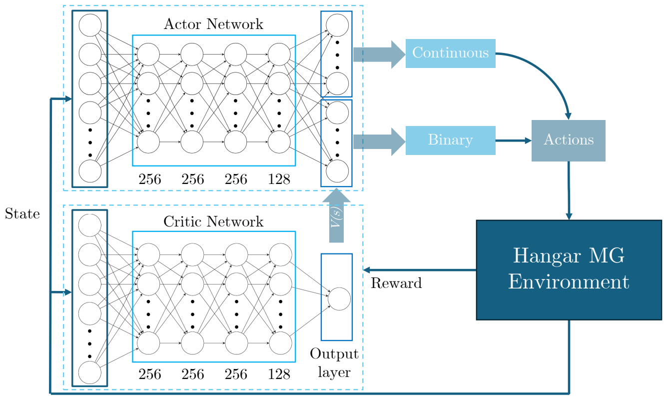

Fig. 4. PPO-based RL architecture for EMS

CPUs, with 2.20 GHz base speed, 512 GB of RAM, and 8 NVIDIA RTX A5000 GPUs, with GPU acceleration enabled via CUDA 11.8.

To highlight the benefits of the proposed physics-based DRL method, Fig. 5 compares its training performance against the classic reward-based DRL approach under the winter scenario. This scenario proved to be the most challenging, since during cold winter days, the EH is active, thus involving a discrete control decision. As shown in Fig. 5, the constrained PPO agent reaches a stable reward region faster and exhibits lower reward variability than the Reward-Based PPO approach. This empirical behavior is attributed to the physics-based action-processing mechanism, which reduces infeasible exploration by restricting the actions applied within the environment to state-dependent feasible operating ranges. Furthermore, the results obtained from the proposed DRL framework, are benchmarked against the MPC-based EMS proposed in [4]. The BESS replacement cost and rated capacity are assumed to be 139 \$/kWh and 20 kWh, respectively, based on [4]. For the case studies, the penalty factors are set to $\lambda_1 = 0.1$ and $\lambda_2 = 0.25$, based on preliminary sensitivity analysis. This showed that the agent's convergence and operating behavior are sensitive to the relative scaling of these coefficients, with the selected values providing stable convergence while maintaining adequate indoor temperature regulation and meaningful BESS operation.

A. Boundary Operating Scenarios

This subsection presents the detailed time-series evaluation of the proposed DRL-based EMS under the two boundary operating scenarios, corresponding to the hottest summer day and the coldest winter day. These scenarios are used to assess the controller performance under high-stress environmental conditions, and to provide a detailed comparison with the MPC benchmark. For these boundary operating scenarios, each seasonal agent is trained for 2,400,000 time steps, corresponding to 25,000 daily training episodes, with a training time of approximately 4 hours. Ten independent training runs are conducted under the stochastic training setup to ensure unbiased performance evaluation, from which the policy with the best validation performance is selected for comparison with the MPC benchmark. In the latter, forecast uncertainty is represented by perturbing the exogenous inputs over the receding prediction horizon, as explained in detail in [4], while the realized daily profiles used for system-state updates are kept fixed for both controllers.

TABLE IV
DAILY COST OF OPERATION OF THE MG

	Summer	Winter
MPC [\$]	34.206	31.780
Constrained PPO [\$]	35.087 ± 0.298	32.497 ± 0.411

The sensitivity of the controller to the training duration was assessed through preliminary training tests using fewer training episodes with the same representative seasonal profiles. These tests showed that indoor temperature regulation remained relatively stable, while BESS operation was more noticeably affected. This behavior is expected because effective BESS operation requires learning longer-term temporal trade-offs among electricity price variations, PV generation, electric demand, and SOC evolution over the complete daily horizon. In contrast, indoor temperature regulation is more directly guided by the imposed physics-based constraints. Thus, the adopted training duration was selected through an empirical sensitivity assessment to improve the consistency of BESS operation while maintaining stable thermal performance and keeping the offline training process computationally tractable.

Fig. 6 and Fig. 7 illustrate the MG active power balance for the hottest summer and coldest winter days, respectively. These and the following figures show the DRL agent control actions, i.e., BESS operation and room temperature control compared with those obtained using the MPC approach. Note that during summer no power export to the grid takes place, whereas in winter, power is exported between 10:30–11:15 am and 12:30–1:15 pm as a result of high solar availability. As depicted in Table IV, the proposed DRL strategy achieves close results compared to the optimal values obtained using the MPC-based EMS benchmark, while requiring substantially lower computation times. Specifically, the average inference times of the DRL framework were 1.219 ms and 1.199 ms for summer and winter, respectively, whereas the corresponding average solve times of the MPC approach were 0.845 s and 4.333 s. This corresponds to solutions being obtained 700 times faster for the summer case and over 3,600 times faster for the winter case. The BESS SOC profiles for the summer and winter scenarios are shown in Fig. 8. Notably, the BESS control achieved by the DRL agent exhibits arbitrage behavior closely resembling that of the MPC benchmark, thus enabling effective cost minimization and constraint satisfaction. Note that the DRL agent tends to take more conservative and incremental dispatch decisions due to its limited perception of the full horizon. Based on a comparative evaluation with respect to the MPC results, Mean Average Errors (MAEs) of 5.252% and 8.903% are observed for summer and winter days, respectively, while the corresponding Root Mean Squared Errors (RMSEs) are 7.371% and 12.507%. For a fair comparison, the initial and final SOC for both the MPC and DRL strategies is assumed to be 50%.

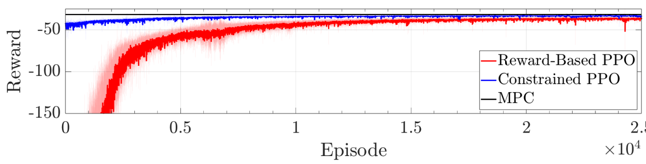

Fig. 5. Cumulative reward over training episodes for the winter scenario.

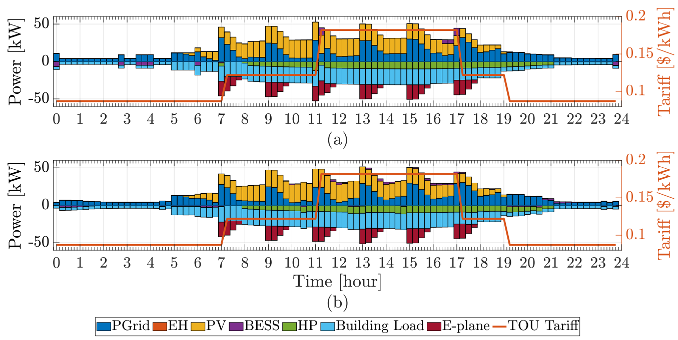

Fig. 6. WWFC hangar MG active power balance for the hottest summer day with (a) MPC approach and (b) proposed RL framework.

The temperature profiles for the summer and winter scenarios are illustrated in Figs. 9 and 10, respectively. Note that the control strategy provided by the DRL agent enables adequate temperature regulation with minimal boundary violations in rooms where temperature control is enforced, thus demonstrating the effectiveness of the proposed approach. Furthermore, the temperature profiles in all rooms closely resemble those obtained using the MPC approach, with two notable exceptions in the summer scenario at 7 am and 11 am, and at 7 am in the winter scenario, due to pre-cooling and pre-heating operations, respectively, which were not captured by the DRL agent. The largest temperature discrepancies are observed in Room 7, where heating is only provided by the EH with a simple ON/OFF thermostat control. Hence, as discussed in Section III-A, the binary EH command is handled through environment-level post-processing of the continuous policy output, whereas the HP actions are further processed through physics-based clipping to enforce state-dependent feasible operating bounds. Thus, unlike the rooms controlled by HPs, Room 7 temperature regulation relies primarily on the reward penalty associated with hangar temperature deviations and on the learned ON/OFF EH switching behavior. Nevertheless, the temperature profiles remain close to those obtained with the MPC benchmark. It is worth mentioning that a more advanced EH device with duty-cycle-based PWM temperature controls may reduce these deviations.

Notably, except for Room 9, whose temperature remains close to the defined limits in both scenarios, the remaining rooms exhibit very low expected occupancy and thus have minimal influence on overall comfort. These rooms correspond to 2 corridors, the washroom and the electrical equipment room. Tables V and VI present a quantitative comparison with the MPC approach, where it can be observed that the proposed DRL framework achieves near-optimal control, with maximum MAE and RMSE values of 0.591°C and 1.583°C, respectively.

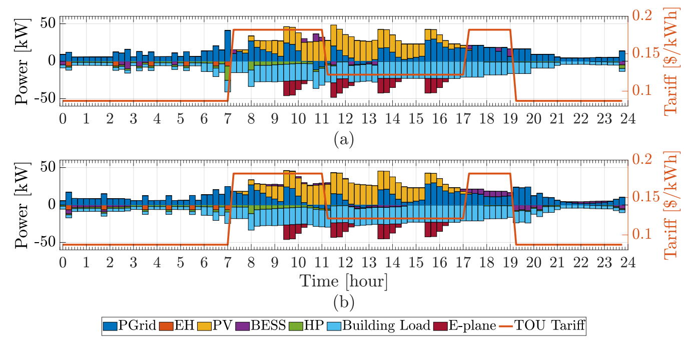

Fig. 7. WWFC hangar MG active power balance for the coldest winter day with (a) MPC approach and (b) proposed RL framework.

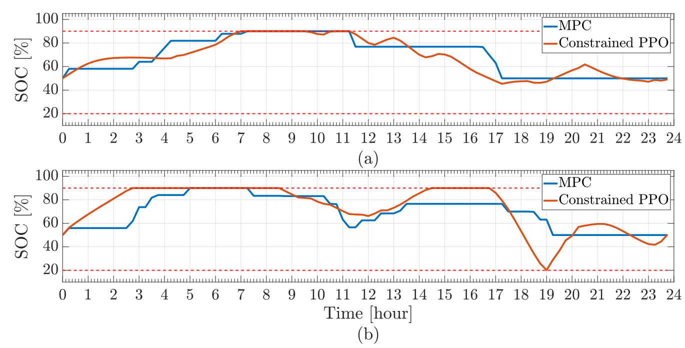

Fig. 8. BESS SOC for (a) hottest summer and (b) coldest winter days.

B. Model Generalization

To first assess the generalization capability of the policies trained under the boundary seasonal scenarios, two unseen days are added for hot summer and cold winter conditions. These days follow comparable daily trends to those observed during training, while still exhibiting differences in magnitude and timing in operating profiles, particularly PV generation and ambient temperature. The obtained results allow to evaluate whether policies trained for specific seasonal boundary cases can retain acceptable performance when applied to operating conditions not seen during training. The corresponding results in Table VII show that the boundary-trained policies retain reasonable performance when evaluated on unseen days with comparable operating trends. Although the SOC MAE values indicate that the battery operation does not fully match the MPC benchmark, the resulting SOC trajectories still capture the main charging and discharging trends, which is relevant because these policies were not specifically optimized for the unseen operating days. More importantly, the temperature results in Table VIII show that the physics-based action constraints embedded in the environment help preserve accurate temperature regulation even when the boundary-trained policies are applied to unseen operating days. Overall, these results demonstrate a degree of transferability of the boundary-trained policies to unseen cases with similar characteristics. However, stronger performance deviations may arise when the unseen operating conditions differ more substantially from the boundary scenarios used during training. Hence, an additional generalization assessment is carried out next by expanding the

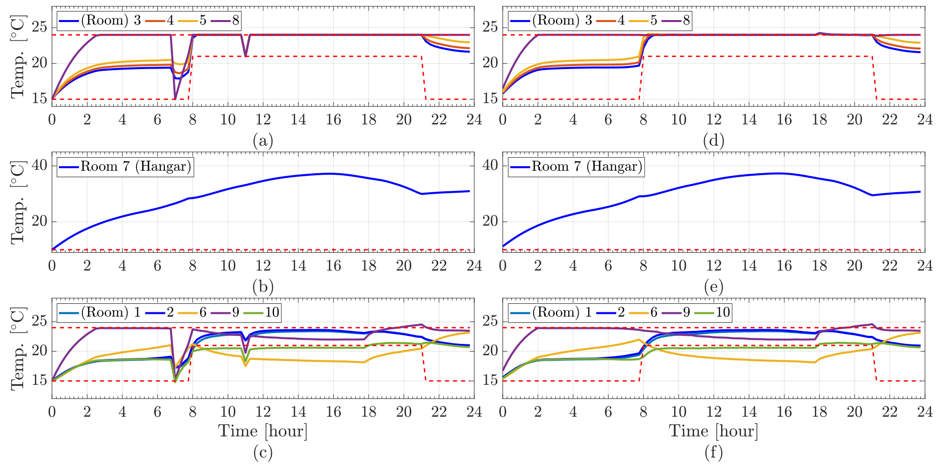

Fig. 9. Indoor temperature profiles in (a) rooms with active temperature control, (b) the planes' hangar, and (c) rooms without temperature control, using the MPC approach; and (d), (e), and (f), respectively, using the proposed RL framework, for the hottest day in Summer.

TABLE V
EVALUATION METRICS FOR TEMPERATURE DEVIATIONS IN SUMMER

Metric	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9	Room 10
MAE [°C]	0.185	0.193	0.218	0.202	0.183	0.286	0.488	0.424	0.427	0.216
RMSE [°C]	0.384	0.410	0.493	0.463	0.430	0.666	0.586	1.383	1.319	0.561

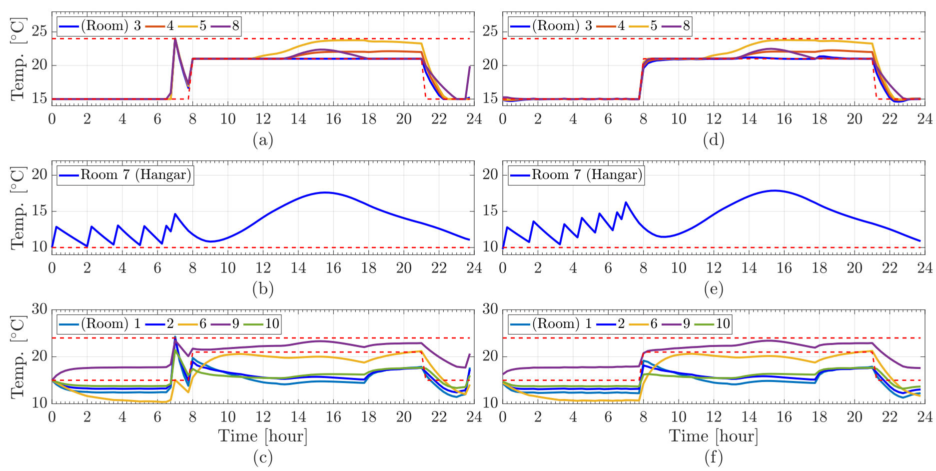

Fig. 10. Indoor temperature profiles in (a) rooms with active temperature control, (b) the planes' hangar, and (c) rooms without temperature control, using the MPC approach; and (d), (e), and (f), respectively, using the proposed RL framework, for the coldest day in Winter.

TABLE VI
EVALUATION METRICS FOR TEMPERATURE DEVIATIONS IN WINTER

Metric	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9	Room 10
MAE [°C]	0.457	0.417	0.326	0.298	0.279	0.312	0.591	0.348	0.311	0.353
RMSE [°C]	1.583	1.437	1.203	1.220	1.217	0.736	0.956	1.315	0.916	1.159

training set to include a broader range of daily profiles.

To further evaluate the generalization capability of the trained seasonal policies, additional daily profiles are considered beyond the boundary operating scenarios previously discussed. For this assessment, the training set is expanded to include 10 days of seasonal variations in PV generation, ambient temperature, building demand, and e-plane charging profiles in summer, fall, winter, and spring days, preserving the training structure adopted in this work for warm and cold days. Each seasonal agent is trained for 4,800,000 time steps, corresponding to approximately 50,000 daily training episodes, resulting in training times in the range of 7 to 8 hours. At the beginning of each episode, one of five diverse daily profiles for the corresponding season is randomly selected, exposing the agent to diverse operating conditions during training. In addition, as previously highlighted, the observation space

TABLE VII
PERFORMANCE OF PROPOSED DRL-BASED EMS VERSUS MPC MODEL FOR A HOT SUMMER DAY AND A COLD WINTER DAY

Scenario	DRL Cost [$]	MPC Cost [$]	Cost Difference [%]	SOC MAE [%]	Avg. DRL Inference Time [ms]	Avg. MPC Solve Time [ms]
Unseen Summer Day	37.663	37.224	+1.177	10.929	1.204	895
Unseen Winter Day	27.257	26.744	+1.920	12.029	1.210	3119

TABLE VIII
TEMPERATURE MAE FOR A HOT SUMMER DAY AND A COLD WINTER DAY

Scenario	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9	Room 10	Mean MAE [°C]
Unseen Summer Day	0.134	0.158	0.167	0.164	0.157	0.241	0.536	0.255	0.268	0.171	0.225
Unseen Winter Day	0.433	0.400	0.344	0.302	0.275	0.362	1.126	0.305	0.279	0.345	0.417

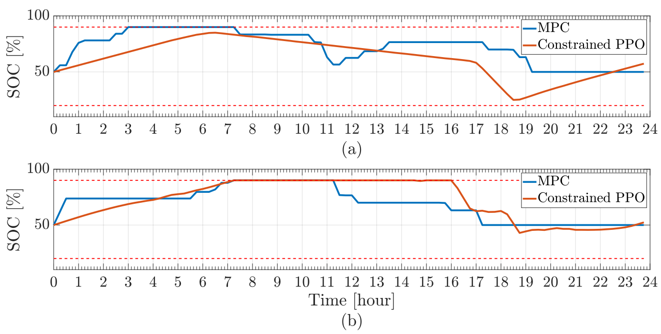

Fig. 11. BESS SOC for (a) Sp^{u} and (b) F^{u} days using the generalized training policy.

includes a 1-hour look-ahead window for the EMS inputs with stochastic perturbations, exposing the agent to multiple forecast realizations around each daily profile. The resulting policies are evaluated on six different test days, including seen-profile and unseen-profile conditions in summer, spring, fall, and winter. The unseen days include cases in summer and winter with operating patterns similar to those used during training, as well as more challenging cases in spring and fall corresponding to completely unseen daily profiles with substantially different operating conditions. The labels for these test days are summarized in Table IX.

The performance of the DRL-based EMS benchmarked against the MPC approach for the various test days is presented in Table X in terms of operating cost, BESS SOC, and solution times. In addition, as previously mentioned, training under cold days is more challenging, which is reflected in the larger cost differences and MAE values observed for scenarios W^{s}, W^{u}, and Sp^{u}. This behavior can be partly attributed to the larger differences in the daily PV profiles observed in cold days, which affect the required battery operation over the daily horizon, whereas the PV profiles for warm days are less variable. The average inference time per control interval of the proposed DRL framework is approximately three orders of magnitude lower than the corresponding average solution of the MPC approach, while maintaining comparable performance in terms of operating cost and SOC tracking.

The corresponding temperature control analysis is illustrated in Table XI, showing that the policy preserves accurate thermal regulation, with low mean temperature MAE values across the test cases. This indicates that feasible and accurate temperature control can be maintained while reducing the dependence on exact model-parameter information during deployment.

TABLE IX
DEFINITION OF GENERALIZATION TEST DAYS

Additional Test Days	Description
S^{s}	Seen-profile summer day
S^{u}	Unseen summer day with comparable operating profiles
F^{u}	Unseen fall day with different operating profiles
W^{s}	Seen-profile winter day
W^{u}	Unseen winter day with comparable operating profiles
Sp^{u}	Unseen spring day with different operating profiles

Figs. 11 and 12 illustrate the SOC and indoor temperature profiles for the unseen spring Sp^{u} and fall F^{u} days, highlighting the consistency of the DRL policy with respect to the MPC benchmark. The overall results show that the DRL-based EMS is able to reproduce the main SOC and thermal-control trends observed in the MPC solution while preserving feasible operation.

The presented results indicate that the proposed training strategy, which combines multiple daily profiles with stochastic perturbations in the look-ahead information, improves the robustness and transferability of the learned policy. Since a single DRL policy is trained to operate across different daily profiles, it is not expected to reproduce the day-specific optimum of the MPC benchmark for every individual case, as shown in Section IV-A. Instead, the policy learns a control strategy that balances performance across different operating conditions and can be applied to unseen profiles. This explains why the BESS operation and cost performance are more affected by differences between individual profiles, as battery scheduling is strongly dependent on the demand profiles, PV generation, and electricity prices. In contrast, the physics-based constraints support effective temperature control and stronger generalization, even under substantial changes in environmental conditions, with low mean MAE values observed for both seen and unseen test days. Overall, the results show that the trained policy can operate under unseen daily conditions, with the strongest generalization performance observed in temperature control.

V. CONCLUSION

In this paper, the implementation of a DRL framework for the optimal operation of a multi-zone TE airport hangar MG was presented, accounting for the thermal needs and resources of the various rooms, as well as e-plane charging based on actual operational data. Constraint satisfaction was

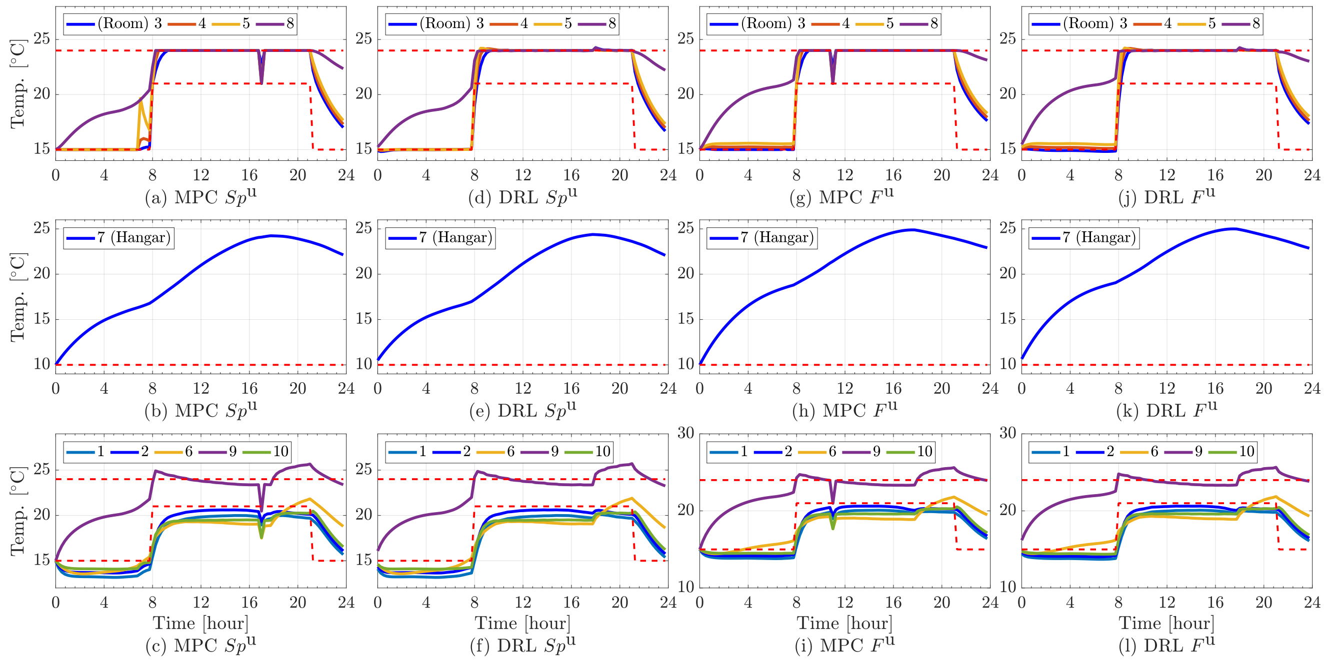

Fig. 12. Indoor temperature profiles for all rooms for unseen spring and fall days.

TABLE X
PERFORMANCE OF PROPOSED DRL-BASED EMS VERSUS MPC MODEL FOR DAYS IN TABLE IX

Day	DRL Cost [$]	MPC Cost [$]	Cost Difference [%]	SOC MAE [%]	Avg. DRL Inference Time [ms]	Avg. MPC Solve Time [ms]
S^{s}	36.671	36.238	+1.193	6.506	1.185	862
S^{u}	39.426	38.988	+1.122	7.104	1.213	810
F^{u}	48.272	47.612	+1.385	7.521	1.343	716
W^{s}	46.608	45.495	+2.446	13.856	1.222	770
W^{u}	47.226	46.114	+2.413	11.318	1.216	830
Sp^{u}	34.796	33.863	+2.756	11.759	1.308	717

TABLE XI
TEMPERATURE MAE FOR DAYS IN TABLE IX

Day	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9	Room 10	Mean MAE [°C]
S^{s}	0.098	0.096	0.099	0.107	0.112	0.136	0.267	0.226	0.247	0.099	0.149
S^{u}	0.133	0.138	0.166	0.162	0.157	0.189	0.370	0.264	0.276	0.147	0.200
F^{u}	0.090	0.091	0.113	0.106	0.089	0.110	0.243	0.185	0.211	0.090	0.133
W^{s}	0.418	0.354	0.203	0.245	0.318	0.323	0.089	0.370	0.319	0.281	0.292
W^{u}	0.428	0.367	0.232	0.313	0.323	0.328	0.092	0.390	0.333	0.307	0.311
Sp^{u}	0.106	0.103	0.104	0.138	0.206	0.114	0.205	0.157	0.188	0.112	0.143

ensured through physics-based dynamic constraints governing the operation of HPs and BESS. Furthermore, a multi-dimensional hybrid action space was considered to account for the operation of an EH within a single-agent control strategy, thus simplifying the overall optimization process. The proposed RL-based EMS was applied to the model of an actual airport MG under deployment at the WWFC in Ontario, Canada, where two boundary scenarios were evaluated, corresponding to the hottest and coldest days in the region. Simulation results were benchmarked against an MPC-based model to assess the performance of the developed DRL framework in terms of multi-room temperature control and BESS operation, using appropriate quantitative metrics, which exhibit minimal variations between the two approaches, highlighting the effectiveness of the proposed DRL-based EMS. Notably, in a practical deployment, DRL approaches offer several advantages over traditional MPC, including the ability to circumvent detailed system modeling, adapt online to changing operating conditions, and achieve substantially faster computation times, which is particularly valuable for real-time MG energy management.

Future work may focus on extending the present simulation-based validation toward experimental deployment and validation of the proposed RL-based EMS at the WWFC MG once it is completed. In this context, real-time measurements would allow the actual thermal dynamics of the airport hangar to be captured more accurately, thereby enhancing the agent's learning capability. This would also support the development of robust policies that can adapt to sensor noise, model uncertainty, and unforeseen operating conditions, potentially through online learning or adaptive control strategies.

Another potential direction for future work is the scalability of the proposed framework for larger airport TE-MGs. The centralized single-agent architecture adopted in this work provides a modular formulation that can accommodate additional zones and DERs within an MG, provided that the policy is properly trained for the expanded system. However, in full-scale airport applications involving multiple terminals, several hangars, additional thermal zones, and a larger number of controllable DERs, the system complexity would increase significantly, particularly due to the need to coordinate multiple buildings and subsystems. In such cases, hierarchical or multi-agent DRL architectures may provide a more scalable alternative.

REFERENCES

[1] I. Hannula, J. B. Menendez, T. Bosoni, A. Bressers, F. Briens, E. Connelly, J. Couse, L. Cret, M. Fajardy, C. Healya, J. Moorhouse, and Y. Om, "The Role of E-fuels in Decarbonising Transport," International Energy Agency, Tech. Rep., 2024.

[2] D. E. Olivares, C. A. Canizares, and M. Kazerani, "A Centralized Energy Management System for Isolated Microgrids," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1864–1875, 2014.

[3] S. Jin and Y. Li, "Analyzing the performance of electricity, heating, and cooling supply nexus in a hybrid energy system of airport under uncertainty," *Energy*, vol. 272, p. 127138, 2023.

[4] P. Verdugo, C. Cañizares, and M. Pirnia, "Modeling and Energy Management of Hangar Thermo-Electrical Microgrid for Electric Plane Charging Considering Multiple Zones and Resources," *Applied Energy*, vol. 379, p. 124951, 2025.

[5] Z. Guo, B. Li, G. Taylor, and X. Zhang, "Infrastructure Planning for Airport Microgrid Integrated with Electric Aircraft and Parking Lot Electric Vehicles," *eTransportation*, vol. 17, p. 100257, 2023.

[6] Z. Liu, Q. Wang, D. Sigler, A. Kotz, K. J. Kelly, M. Lunacek, C. Phillips, and V. Garikapati, "Data-Driven Simulation-Based Planning for Electric Airport Shuttle Systems: A Real-World Case Study," *Applied Energy*, vol. 332, p. 120483, 2023.

[7] B. Hou, S. Bose, L. Marla, and K. Haran, "Impact of Aviation Electrification on Airports: Flight Scheduling and Charging," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 3, pp. 2342–2354, 2024.

[8] P. Ollas, S. G. Sigarchian, H. Alfredsson, J. Leijon, J. S. Döhler, C. Aalhuizen, T. Thiringer, and K. Thomas, "Evaluating the Role of Solar Photovoltaic and Battery Storage in Supporting Electric Aviation and Vehicle Infrastructure at Visby Airport," *Applied Energy*, vol. 352, p. 121946, 2023.

[9] H. Zhao, Y. Xiang, Y. Shen, Y. Guo, P. Xue, W. Sun, H. Cai, C. Gu, and J. Liu, "Resilience Assessment of Hydrogen-Integrated Energy System for Airport Electrification," *IEEE Transactions on Industry Applications*, vol. 58, no. 2, pp. 2812–2824, 2022.

[10] W. Violante, C. A. Cañizares, M. A. Trovato, and G. Forte, "An Energy Management System for Isolated Microgrids with Thermal Energy Resources," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 2880–2891, 2020.

[11] Z. Li, L. Wu, Y. Xu, S. Moazeni, and Z. Tang, "Multi-Stage Real-Time Operation of a Multi-Energy Microgrid With Electrical and Thermal Energy Storage Assets: A Data-Driven MPC-ADP Approach," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 213–226, 2022.

[12] G. Ceusters, R. C. Rodríguez, A. B. García, R. Franke, G. Deconinck, L. Helsen, A. Nowé, M. Messagie, and L. R. Camargo, "Model-Predictive Control and Reinforcement Learning in Multi-Energy System Case Studies," *Applied Energy*, vol. 303, p. 117634, 2021.

[13] H. H. Goh, Y. Huang, C. S. Lim, D. Zhang, H. Liu, W. Dai, T. A. Kurniawan, and S. Rahman, "An Assessment of Multistage Reward Function Design for Deep Reinforcement Learning-Based Microgrid Energy Management," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4300–4311, 2022.

[14] J. Arroyo, C. Manna, F. Spiessens, and L. Helsen, "Reinforced Model Predictive Control (RL-MPC) for Building Energy Management," *Applied Energy*, vol. 309, p. 118346, 2022.

[15] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[17] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-Line Building Energy Optimization Using Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698–3708, 2019.

[18] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 2020.

[19] R. Lu, S. H. Hong, and M. Yu, "Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6629–6639, 2019.

[20] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-Free Real-Time Autonomous Control for a Residential Multi-Energy System Using Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068–3082, 2020.

[21] Y. Li, S. He, Y. Li, Y. Shi, and Z. Zeng, "Federated Multiagent Deep Reinforcement Learning Approach via Physics-Informed Reward for Multimicrogrid Energy Management," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 5902–5914, 2024.

[22] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao, "Multi-Agent Safe Policy Learning for Power Management of Networked Microgrids," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1048–1062, 2021.

[23] R. Shen, S. Zhong, X. Wen, Q. An, R. Zheng, Y. Li, and J. Zhao, "Multi-Agent Deep Reinforcement Learning Optimization Framework for Building Energy System with Renewable Energy," *Applied Energy*, vol. 312, p. 118724, 2022.

[24] Y. Li, J. Gao, Y. Li, C. Chen, S. Li, M. Shahidehpour, and Z. Chen, "Physical Informed-Inspired Deep Reinforcement Learning Based Bi-Level Programming for Microgrid Scheduling," *IEEE Transactions on Industry Applications*, vol. 61, no. 1, pp. 1488–1500, 2025.

[25] Y. Zhang, Z. Mei, X. Wu, H. Jiang, J. Zhang, and W. Gao, "Two-Step Diffusion Policy Deep Reinforcement Learning Method for Low-Carbon Multi-Energy Microgrid Energy Management," *IEEE Transactions on Smart Grid*, vol. 15, no. 5, pp. 4576–4588, 2024.

[26] A. S. Omar and R. El-Shatshat, "Energy Management of Multi-Energy Communities: A Hierarchical MIQP-Constrained Deep Reinforcement Learning Approach," *IEEE Transactions on Sustainable Energy*, vol. 16, no. 3, pp. 2236–2250, 2025.

[27] X. Wang, P. Wang, R. Huang, X. Zhu, J. Arroyo, and N. Li, "Safe Deep Reinforcement Learning for Building Energy Management," *Applied Energy*, vol. 377, p. 124328, 2025.

[28] Y. Zhang, Y. Ren, Z. Liu, H. Li, H. Jiang, Y. Xue, J. Ou, R. Hu, J. Zhang, and D. W. Gao, "Federated Deep Reinforcement Learning for Varying-Scale Multi-Energy Microgrids Energy Management Considering Comprehensive Security," *Applied Energy*, vol. 380, p. 125072, 2025.

[29] G. Ceusters, L. R. Camargo, R. Franke, A. Nowé, and M. Messagie, "Safe Reinforcement Learning for Multi-Energy Management Systems with Known Constraint Functions," *Energy and AI*, vol. 12, p. 100227, 2023.

[30] Y. Ye, H. Wang, P. Chen, Y. Tang, and G. Strbac, "Safe Deep Reinforcement Learning for Microgrid Energy Management in Distribution Networks With Leveraged Spatial–Temporal Perception," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3759–3775, 2023.

[31] X. Ding, A. Cerpa, and W. Du, "Multi-Zone HVAC Control With Model-Based Deep Reinforcement Learning," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 4408–4426, 2025.

[32] W. Mendieta and C. A. Cañizares, "Primary Frequency Control in Isolated Microgrids Using Thermostatically Controllable Loads," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 93–105, 2021.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[34] B. Xu, J. Zhao, T. Zheng, E. Litvinov, and D. S. Kirschen, "Factoring the Cycle Aging Cost of Batteries Participating in Electricity Markets," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2248–2259, 2018.

[35] "Ventilation and Acceptable Indoor Air Quality," American Society of Heating, Refrigerating and Air-Conditioning Engineers, Standard, 2022.

[36] "Velis Electro Pilot's Operating Handbook," Pipistrel, Handbook, 2020.

[37] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," 2019.