Optimal Operation of Hydrogen-based Multi-Energy System via a Self-Predictive Deep Reinforcement Learning Approach

Zhenyu Pu, Yu Yang, *Member, IEEE,* Lun Yang, *Member, IEEE,* Qing-Shan Jia, *Senior Member, IEEE,* Xiaohong Guan, *Life Fellow, IEEE,* Costas J. Spanos, *Fellow, IEEE*

***Abstract*—Hydrogen-based multi-energy systems (MES) have attracted growing attention in the context of low-carbon and sustainable energy system transition due to their significant potential in improving overall energy efficiency and facilitating renewable utilization. However, it requires to address the optimal operation of the system under complex multi-energy flow couplings, inherently nonlinear and spatially-temporally coupled system dynamics and multiple sources of uncertainties. To address this challenge, this paper first proposes a comprehensive operational optimization model for hydrogen-based MES that fully captures the nonlinear and coupled thermal, electrochemical and pressure dynamics of electrolyzers, fuel cells, and hydrogen tanks. Then, a novel self-predictive deep reinforcement learning method (SP-DRL) is proposed to enable learning-based optimal operation of hydrogen-based MES. Particularly, we learn a latent dynamic model of the system and incorporate it into the DRL framework to address the limitations of conventional DRL approaches for complex systems. Extensive case studies based on real-world data demonstrate that the proposed method significantly improves conventional DRL approaches. In particular, the SP-DRL method significantly enhances the convergence speed, training stability and policy performance. Moreover, by the proposed SP-DRL method, the external energy purchase costs of hydrogen-based MES is reduced by about 11-32 % across all tested scenarios.**

***Index Terms*—Multi-energy systems (MES), Hydrogen energy, Self-predictive Deep reinforcement learning (SP-DRL), Operational optimization of complex energy systems**

I. Introduction

DRIVEN by the dual pressure of environmental deterioration and fossil fuel depletion, global energy systems are facing a profound low-carbon and sustainable transition [1]. Multi-energy systems (MES) that integrate diverse energy generation, conversion and storage devices to enable coordinated multi-energy (i.e., electricity, heating and cooling) production and consumption have been recognized as an important means to advance the energy transition [2]. This is largely due to their significant potential in enhancing overall energy efficiency and facilitating the consumption of volatile and nondispatchable renewable energy (i.e., wind and solar power). In this context, hydrogen-based MES have attracted growing attention to enable low-carbon or zero-carbon energy systems due to the many unique advantages of hydrogen as an energy carrier [3–5]. First and foremost, hydrogen has a wide range of sources, including industrial by-products, fossil-fuel based reforming and water electrolysis, making it particularly suitable as a secondary energy carrier. In addition, its relatively low self-discharge rate makes it well suited for large-scale and long-duration energy storage, enabling cross-seasonal and cross-regional shifting of intermittent renewable energy. Moreover, hydrogen shows strong multi-energy integration capability as it can be flexibly converted into different energy forms, including electricity, heating and cooling. Owing to these advantages, hydrogen has been recognized as an important energy carrier to achieve China's carbon neutrality target [6].

While hydrogen-based MES are promising to enable zero-carbon or renewable-powered energy systems [7, 8], the effective operation of such systems remains a key issue. Specifically, an intelligent controller that can dynamically regulate the operation of diverse energy generation, conversion and storage devices to ensure consistent multi-energy supply-demand balance while achieving the low-carbon and energy-efficient target is essential. This issue is challenging due to the following aspects. First, there exist tight temporal and spatial couplings across the diverse energy generation, conversion and storage devices. Besides, many of the energy devices show intrinsic nonlinear and coupled operating characteristics due to the multi-energy conversion process. Moreover, there exist multiple sources of uncertainties, arising from both renewable energy supply and the multi-energy demands.

In recent years, considerable efforts have been devoted to addressing this challenge. For example, Liu et al. [9] studied the optimal operation of hydrogen-based MES considering the long-term operational economic benefits based on deterministic programming. The works [10, 11] further studied the optimal operation of the system considering multiple uncertainties based on stochastic programming (SP). Fang et al. [12] studied the operation of multiple hydrogen-based MES considering energy transaction based on model predictive control (MPC) methods. Chen et al. [13] investigated the optimal operation of hydrogen-based MES in a building community considering building thermal load flexibility by a two-layer MPC approach. Sui el al. [14] studied the short-term scheduling strategy of integrated electrichydrogen-Thermal systems that considers the nonlinear and coupled operating dynamics of hydrogen system by a mixed-integer second-order conic programming (MISOCP) approach. The above works are mainly model-based optimization approaches that rely on a computationally tractable mathematical model to capture the system dynamics. They obtain stage-wise operational decisions through solving constrained optimization problems of minimizing energy or operational cost while considering the device operating limits and supply-demand balance constraints. While these methods are advantageous in capturing the physical operating limits and multi-energy supply-demand balances constraints, their intractability and computational efficiency largely depend on the complexity of mathematical models.

Driven by the proliferation of artificial intelligence (AI), deep reinforcement learning (DRL) recently have raised extensive attention in energy system optimization due to its capability of handling complex system dynamics and multiple sources of uncertainties [15]. Specifically, DRL methods do not rely on computationally efficient mathematical models of energy systems, and they learn operational policies from the trial-and-error of system operation. A growing body of work has focused on DRL-based approaches for MES. For example, Franzoso et al. [16] investigated the direct application of DRL to MES and showed that it outperforms conventional rule-based control policies. Dolatabali et al. [17] studied the combination of DRL with CNN and LSTM for hydrogen-based MES under uncertainty. Specifically, the method takes high-resolution sky images and numerical meteorological data as inputs to predict solar radiation, which was then incorporated into the DRL framework to support informed decision-making. Zhao et al. [18] studied DRL-based energy management for household MES and demonstrated that it outperforms prediction-based MPC under the multiple uncertainties. While the above DRL methods depend on raw data to implicitly capture the uncertainties, many recent works have incorporated explicit uncertainty modeling into DRL framework to better address the uncertainties. For example, Zhang et al. [19] studied the combination of deep learning based interval predictions with DRL for hydrogen-based MES and demonstrated that it significantly improved operational performance. Zhang et al. [20] combined Bayesian probability distribution modeling with DRL for multi-energy microgrids, and demonstrated that it achieved near-optimal operating policy.

Despite the above research progress, the optimal operation of hydrogen-based MES still has not yet been well addressed. On one hand, existing works mainly relied on simplified linear models to characterize the operating characteristics of hydrogen system, such as the electrolyzers, hydrogen tanks and fuel cells [9, 11–13, 21]. In such works, the energy conversion efficiencies of electrolyzers and fuel cells were assumed constant over their entire operating ranges. Some works used piece-wise linear functions to approximately capture the varying operating efficiency of fuel cells [10, 16]. Both the linear and piece-wise linear formulations are computationally attractive for model-based optimization, however they are insufficient to capture the complex nonlinear operating characteristics of hydrogen systems. The computed operating policies may deviate far from the actual optimal point. Some works sufficiently considered the nonlinear operating characteristics of hydrogen system but face intensive computational complexity [14, 22]. Moreover, most existing studies neglect the thermal, electrochemical, and pressure dynamics of electrolyzers and fuel cells due to computational complexity. However, these dynamics are closely coupled with system operation and have a substantial impact on device efficiency and overall system performance [23–25]. It is therefore important to fully consider the nonlinear and coupled operating dynamics to ensure effective, reliable and energy-efficient operation of hydrogen-based MES. On the other hand, while DRL are promising to handle the complex system dynamics and multiple sources of uncertainties [16, 17], existing works are mainly direct applications of conventional DRL approaches. These methods fully rely on operational data to implicitly capture the complex system dynamics in the DRL pipeline. As a result, they often show limited performance in convergence rate, sampling efficiency, training stability and policy performance for complex energy systems. In recent years, state representation learning and latent dynamic model learning have attracted growing interest to enhance the performance of DRL [26, 27]. The core idea is to learn compact and informative state representations or latent dynamic models of environments, thus facilitating effective DRL learning. State representation learning originates from vision-based DRL where the observations are high-dimensional images and required to be compressed into low-dimensional and compact representations to enable effective DRL learning [28, 29]. Latent dynamics learning focuses on learning a transition model of the environment or system in a latent space, which can then be used to guide and improve DRL policy learning [30, 31]. These works proposed promising directions to develop advanced DRL to addressing the operational complexity of energy systems. However, their practical implementation remains largely unexplored. Challenging issues include how a dynamic model can be learned and how it can be incorporated into the DRL pipeline to achieve effective policy learning.

Motivated by the above researches, this paper investigates the operational optimization of a hydrogen-based MES under uncertainty. Our main contributions are as follows.

- We proposed a comprehensive operational optimization model for hydrogen-based MES that fully capture the nonlinear and coupled thermal, electrochemical and pressure dynamics of electrolyzers, fuel cells and hydrogen tank to enable a reliable operation of the system.
- We further developed a self-predictive deep reinforcement learning (SP-DRL) approach to enable effective learning-based operation of hydrogen-based MES under uncertainty. Particularly, the method proposes to learn a latent dynamic model of the system and incorporate it into the DRL framework to address the limitations of conventional DRL approaches for complex systems.
- Extensive case studies based on real-world data demonstrated the proposed method significantly improves conventional DRL approaches in convergence speed, sampling efficiency, training stability and policy performance. Particularly, the external energy purchase cost of hydrogen-based MES is reduced by about 11-32% compared with existing DRL across all tested scenarios.

The rest of the paper is as follows. Section II introduces the operational optimization model for hydrogen-based MES. Section III introduces the SR-DRL method for hydrogen-based MES. Section IV presents and analyze case studies. Section V concludes this paper.

II. Problem formulation

A. Hydrogen-based MES

We consider a hydrogen-based MES deployed within a building community as illustrated in Fig. 1. The MES integrates diverse energy generation, conversion and storage

devices to enable coordinated multi-energy (electricity, heating and cooling) supply of the community. At the center of the MES is a hydrogen system composed of an electrolyzer, a compressor unit, a hydrogen tank, and a fuel cell. This system manages hydrogen production, storage, utilization and more importantly enable the multi-energy integration to enhance system operational flexibility. Specifically, the electrolyzer can convert electricity into hydrogen through water electrolysis, and the fuel cell can convert hydrogen back into electricity and heat. A heat recovery unit is coupled with the fuel cell for capturing the heat, which then can be used to heat water in the hot water tank or converted into cooling by the absorption chiller (AC). In addition to hydrogen tank, the MES involves an electrical storage, hot and chilled water tanks as energy storage. The hot and chilled water tanks serve as thermal and cooling energy storage to satisfy diverse forms of heating and cooling demands. A solar thermal collector is also deployed to harvest solar power for producing heat water.

This paper focuses on the optimal operation of the hydrogen-based MES to achieve multi-energy supply-demand balance with minimal external energy purchase cost from the utility grid and hydrogen market. We consider a discrete framework with a decision interval of Δt and each day is equally divided into the set of time slots T. A bottom-up modeling framework is adopted, where the mathematical formulations of individual components are first presented and then the system-level optimization objective is introduced. We use the alphabets P, g, q, v to represent electricity, heating, cooling and hydrogen energy flows throughout the paper.

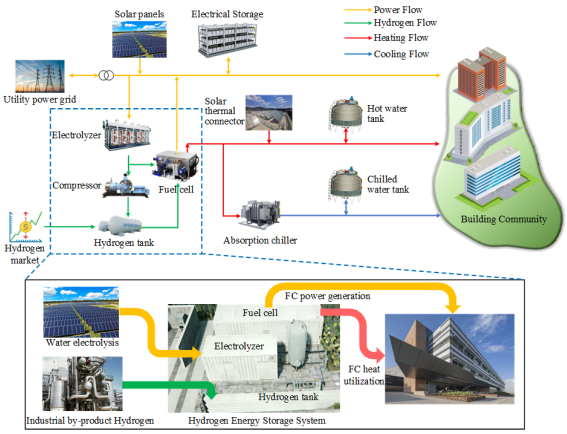

Fig. 1: Hydrogen-based MES in a building community

B. Energy Storage Devices

Electrical storage system (ESS): The ESS is lithium batteries, whose operating dynamics and physical limits over the scheduling horizon T are modeled as

$$S^{\text{ess}}_{t+\Delta t} = S^{\text{ess}}_t + \left(P^{\text{ess,ch}}_t \eta^{\text{ess,ch}} - P^{\text{ess,dis}}_t / \eta^{\text{ess,dis}}\right)\Delta t, \tag{1a}$$

$$0 \le P^{\text{ess,ch}}_t, P^{\text{ess,dis}}_t \le P^{\text{ess}}_{\max}, \tag{1b}$$

$$P^{\text{ess,ch}}_t, P^{\text{ess,dis}}_t = 0, \tag{1c}$$

$$S^{\text{ess}}_{\min} \le S^{\text{ess}}_t \le S^{\text{ess}}_{\max}, \quad t \in T. \tag{1d}$$

where t denotes the time. S^{ess}_t [kWh] denotes the stored electricity. $P^{\text{ess,ch}}_t, P^{\text{ess,dis}}_t$ [kW] are scheduled charging and discharging power, with $\eta^{\text{ess,ch}}, \eta^{\text{ces,dis}}$ denoting charging and discharging efficiencies. Constraints (1a)–(1d) are stored energy dynamics, power limits, non-simultaneous charging and discharging operation, and the permissible operating range of storage capacity.

Hot water tank: The hot water tank serves as thermal energy storage (TES) to meet the diverse forms of heating demands of building community. Its operating dynamics and constraints are captured by [10, 11]

$$S^{\text{tes}}_{t+\Delta t} = S^{\text{tes}}_t + \left(g^{\text{tes,ch}}_t \eta^{\text{tes,ch}} - g^{\text{tes,dis}}_t / \eta^{\text{tes,dis}}\right)\Delta t, \tag{2a}$$

$$0 \le g^{\text{tes,ch}}_t, g^{\text{tes,dis}}_t \le g^{\text{tes}}_{\max}, \tag{2b}$$

$$g^{\text{tes,ch}}_t, g^{\text{tes,dis}}_t = 0, \tag{2c}$$

$$S^{\text{tes}}_{\min} \le S^{\text{tes}}_t \le S^{\text{tes}}_{\max}, \quad t \in T. \tag{2d}$$

where S^{tes}_t [kWh] represents the stored heating energy in the tank. $g^{\text{tes,ch}}_t, g^{\text{tes,dis}}_t$ [kW] are injected and released heat power, with $\eta^{\text{tes,ch}}$ and $\eta^{\text{tes,dis}}$ denoting the corresponding efficiencies.

Chilled Water Tank: The chilled water tank is used as cooling energy storage (CES) for satisfying the diverse cooling demand of building community. Its operating dynamics and constraints are captured by [10, 11]

$$S^{\text{ces}}_{t+\Delta t} = S^{\text{ces}}_t + \left(q^{\text{ces,ch}}_t \eta^{\text{ces,ch}} - q^{\text{ces,dis}}_t / \eta^{\text{tes,dis}}\right)\Delta t, \tag{3a}$$

$$0 \le q^{\text{ces,ch}}_t, q^{\text{ces,dis}}_t \le q^{\text{ces}}_{\max}, \tag{3b}$$

$$q^{\text{ces,ch}}_t q^{\text{ces,dis}}_t = 0, \tag{3c}$$

$$S^{\text{ces}}_{\min} \le S^{\text{ces}}_t \le S^{\text{ces}}_{\max}, \quad t \in T. \tag{3d}$$

where S^{ces}_t [kWh] denotes the stored cooling energy. $q^{\text{ces,ch}}_t, q^{\text{ces,dis}}_t$ [kW] are injected and released cooling power, with $\eta^{\text{ces,ch}}, \eta^{\text{ces,dis}}$ denoting the efficiencies.

C. Hydrogen System

We adopt the gray-box model proposed in [23, 24] to capture the nonlinear and coupled operating dynamics of a hydrogen system formed by a Polymer Electrolyte Membrane (PEM) electrolyzer, a Proton Exchange Membrane Fuel Cell (PEMFC) and a high-pressure gas hydrogen storage tank.

Electrolyzer: The electrolyzer is a self-contained system consisting of four stacks, each of which has an equal nominal rated power. The stacks can be run at an overloaded power for certain time period. To ensure reliable operation, stack temperature is required to be maintained below a prescribed upper limit. A cooling device with a programmed control logic is often coupled with electrolyzer for thermal management. For example, the cooling device is programmed to on and off when stack temperature exceeds 50°C and drops below 40°C for maintaining a target stack temperature of 60°C [24]. The control inputs of the electrolyzer are injected DC power P^{ely}_t within the permissible range of $[P^{\text{ely}}_{\min}, P^{\text{ely}}_{\max}]$ kW with a nominal power of $P^{\text{ely}}_{\text{nom}}$ kW. The energy conversion efficiency of electrolyzer are influenced by the injected DC power and stack temperature, exhibiting a nonlinear characteristic, which can be captured by the piece-wise affine function [24]

$$v^{\text{ely}}_t = \begin{cases} z_1 T^{\text{ely}}_t + z_0 + z_{\text{low}}(P^{\text{ely}}_t - P^{\text{ely}}_{\text{nom}}) & \text{if } P^{\text{ely}}_{\min} < P^{\text{ely}}_t \le P^{\text{ely}}_{\text{nom}}, \\ z_1 T^{\text{ely}}_t + z_0 + z_{\text{high}}(P^{\text{ely}}_t - P^{\text{ely}}_{\text{nom}}) & \text{if } P^{\text{ely}}_{\text{nom}} < P^{\text{ely}}_t \le P^{\text{ely}}_{\max}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where v_t^{ely} [m^3/h] denotes the hydrogen production rate. P_t^{ely} and T_t^{ely} represent the injected DC power and stack temperature. The power-dependent operating efficiency of the electrolyzer is characterized by the model parameters z_{low} and z_{high}. In general, electrolyzers exhibit higher energy conversion efficiency at lower DC power levels, and z_{low} is typically greater than z_{high}.

The stack temperature T_t^{ely} of the electrolyzer is governed by the internal heat generation rate $\dot{Q}_{\text{gen}}$ of hydrogen production, heat dissipation rate $\dot{Q}_{\text{loss}}$ and the cooling power $\dot{Q}_{\text{cool}}$ supplied by the cooling device. The stack temperature dynamics can be capture by a lumped-parameter model [24]

$$\frac{dT_t^{\text{ely}}}{dt} = K_1\dot{Q}_{\text{gen}} - K_2\dot{Q}_{\text{loss}} - K_3\dot{Q}_{\text{cool}} \tag{5}$$

Further, the following discrete approximate model can be adopted to capture the stack temperature dynamics of electrolyzer [24]

$$T_t^{\text{ely}} = j_1 T_{t-\Delta t}^{\text{ely}} + j_2 P_{t-\Delta t}^{\text{ely}} + j_0, \quad \forall t \in T. \tag{6}$$

where the effects of heat generation, loss and cooling devices are encapsulated in model parameters j_0, j_1, j_2. Particularly, the injected DC power P_t^{ely} is involved to implicitly capture the heat generation effects. To ensure reliable operation of electrolyzer, the stack temperature is required to be maintained below a upper limit $T_{\max}^{\text{ely}}$ (often 70°C)

$$0 \le T_t^{\text{ely}} \le T_{\max}^{\text{ely}}, \quad \forall t \in T. \tag{7}$$

Though the electrolyzer has a nominal power, it can be overloaded for certain time period without causing any adverse effects. This behavior can be captured by an overload counter for recording the number of time slots that the stack current overloads [24]. Specifically, the first stack of the electrolyzer can be used as an overload indicator as the injected DC power is often allocated to it first. The corresponding allocated DC power $P_{t,1}^{\text{ely}}$ can be approximately captured by the following PWL function [24]

$$P_{t,1}^{\text{ely}} = \begin{cases} P_t^{\text{ely}} & \text{if } P_{\min}^{\text{ely}} < P_t^{\text{ely}} \le P_{\text{nom}}^{\text{ely}}/4, \\ P_{\text{nom}}^{\text{ely}}/4 & \text{if } P_{\text{nom}}^{\text{ely}}/4 < P_t^{\text{ely}} \le P_{\text{nom}}^{\text{ely}}, \\ P_t^{\text{ely}}/4 & \text{if } P_{\text{nom}}^{\text{ely}} < P_t^{\text{ely}} \le P_{\max}^{\text{ely}}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The resulting stack current can be captured by

$$i_t^{\text{ely}} = \begin{cases} h_1 T_t^{\text{ely}} + h_0 + h_{\text{low}}(P_{t,1}^{\text{ely}} - P_{\text{nom},1}^{\text{ely}}) & \text{if } P_{\min}^{\text{ely}} < P_{t,1}^{\text{ely}} \le P_{\text{nom},1}^{\text{ely}}, \\ h_1 T_t^{\text{ely}} + h_0 + h_{\text{high}}(P_{t,1}^{\text{ely}} - P_{\text{nom},1}^{\text{ely}}) & \text{if } P_{\text{nom},1}^{\text{ely}} < P_{t,1}^{\text{ely}} \le P_{\max}^{\text{ely}}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where $P_{\text{nom},1}^{\text{ely}}$ [kW] denotes the nominal stack power. h_0, h_1, $h_{\text{low}}, h_{\text{high}}$ are model parameters, capturing effects of stack temperature and DC power on stack current. A counter that tracks overload behavior of the first stack can be defined as

$$C_t^{\text{ely}} = \max\left\{0, C_{t-\Delta t}^{\text{ely}} + (i_t^{\text{ely}} - i_{\text{nom}}^{\text{ely}}) \cdot \Delta t\right\}, \tag{10a}$$

$$C_t^{\text{ely}} \le C_{\max}^{\text{ely}}, \quad \forall t \in T. \tag{10b}$$

where $i_{\text{nom}}^{\text{ely}}$ denotes the nominal stack current. $C^{\max}$ denotes the upper limits of overload which is determined by the permissible overload time period.

Hydrogen tank: Hydrogen produced by the electrolyzer can be cleaned, purified and compressed by the compressor and then stored in the hydrogen tank. When needed, the stored hydrogen can be consumed by fuel cell to generate electricity and heat. The reliable operation of hydrogen tank should account for the tank capacity, tank temperature and pressure limits, which can be modeled as [23]

$$S_t^{\text{hss}} = S_{t-\Delta t}^{\text{hss}} + \left(\alpha \cdot v_t^{\text{ely}} - v_t^{\text{fc}} + v_t^{\text{buy}}\right)\rho_0 \Delta t, \tag{11a}$$

$$-v_{\max}^{\text{buy}} \le v_t^{\text{buy}} \le v_{\max}^{\text{buy}}, \tag{11b}$$

$$S_{\min}^{\text{hss}} \le S_t^{\text{hss}} \le S_{\max}^{\text{hss}}, \tag{11c}$$

$$p_{\min}^{\text{tank}} \le p_t^{\text{tank}} \le p_{\max}^{\text{tank}}, \tag{11d}$$

$$T_t^{\text{tank}} = g_0 T_{t-\Delta t}^{\text{tank}} + g_1 T_{t-\Delta t}^{\text{amb}}, \tag{11e}$$

$$p_t^{\text{tank}} = (b_0 + b_1 T_t^{\text{tank}}) \cdot S_t^{\text{hss}} / V^{\text{tank}}, \quad \forall t \in T. \tag{11f}$$

where S_t^{tank} [kg] represents the stored hydrogen, ρ_0 [kg/m^3] denotes the hydrogen density, and V^{tank} denotes the volume of hydrogen tank under standard atmospheric pressure. $T_t^{\text{tank}}, T_t^{\text{amb}}$ and p_t^{tank} represent the tank temperate, ambient temperature and tank pressure. Constraints (11a) characterize the stored hydrogen in the tank, which is determined by production rate of electrolyzer v_t^{ely}, consumption rate of fuel cell v_t^{fc}, and the net purchase in hydrogen market v_t^{buy}. The small non-negative constant $\alpha \in [0,1]$ is to capture the loss of produced hydrogen while through compressor. Constraints. (11b) model the transaction limits in hydrogen market. Constraints (11c) capture lower and upper limits of hydrogen that can be stored in the tank. In addition to those limits, the reliable operation of hydrogen tank should account for its pressure limits as modeled in (11d), which is determined by the stored hydrogen and tank temperature as modeled in (11e) and (11f).

Fuel cell: We consider a Proton Exchange Membrane Fuel Cell (PEMFC) comprising two stacks with a total maximum power of $P_{\max}^{\text{fc}}$ in total. Fuel cell first converts hydrogen v_t^{fc} into stack current i_t^{fc}, and then DC power P_t^{fc}. This process exhibits nonlinear characteristics and can be modeled as [23]

$$v_t^{\text{fc}} = c \cdot i_t^{\text{fc}}, \tag{12a}$$

$$i_t^{\text{fc}} = \begin{cases} s_1 P_t^{\text{fc}}, & \text{if } P_{\min}^{\text{fc}} < P_t^{\text{fc}} \le P_{\text{bp}}^{\text{fc}}, \\ s_2(P_t^{\text{fc}} - P_{\text{bp}}^{\text{fc}}) + i_{\text{bp}}^{\text{fc}}, & \text{if } P_{\text{bp}}^{\text{fc}} < P_t^{\text{fc}} \le P_{\max}^{\text{fc}}, \\ 0, & \text{otherwise,} \end{cases} \tag{12b}$$

$$P_{\min}^{\text{fc}} \le P_t^{\text{fc}} \le P_{\max}^{\text{fc}}, \quad \forall t \in T. \tag{12c}$$

where v_t^{fc} [m^3/h] denotes the rate of injected hydrogen into the fuel cell. c is an experimental constant accounting for purging and Faradaic losses. The conversion of DC power to stack current shows a nonlinear characteristics as captured by the PWL function (12b) with $(P_{\text{bp}}^{\text{fc}}, i_{\text{bp}}^{\text{fc}})$ denoting the breakpoints and s_1, s_2 as energy conversion efficiencies. Constraints (12c) characterize the range of DC power output of the fuel cell.

While generating electricity, fuel cell also produces heat which can be captured by heat recover unit [32]

$$g_t^{\text{fc}} = \eta_{\text{rec}}^{\text{fc}} \cdot (1 - \eta^{\text{fc}}) / \eta^{\text{fc}} \cdot P_t^{\text{fc}}, \forall t \in T. \tag{13}$$

where η_{rec} denotes the heat recovery efficiency and η^{fc} represents the heat conversion efficiency of the fuel cell.

D. Energy Conversion Devices

Absorption Chiller (AC): AC can convert heat into cooling energy. The produced cooling power is determined by the injected heat g_t^{ac} and energy conversion efficiency η^{ac} of AC, which can be modeled as

$$\begin{aligned} q_t^{\text{ac}} &= g_t^{\text{ac}} \cdot \eta^{\text{ac}}, \\ g_t^{\text{ac}} &\le g_{\max}^{\text{ac}}, \quad t \in T. \end{aligned} \tag{14}$$

where g_t^{ac} denotes the injected heat to AC. q_t^{ac} is produced cooling energy. $g_{\max}^{\text{ac}}$ captures the operating limits of the AC.

Photovoltaic power (PV): Solar power generation is determined by the incident solar irradiance and panel area, which is typically modeled as

$$P_t^{\text{solar}} = \eta^{\text{pv}} \cdot A_{\text{pv}} \cdot Q_t^{\text{rad}}, \quad \forall t \in T. \tag{15}$$

where P_t^{solar} [kW] denotes the solar generation, jointly determined by the PV conversion efficiency η^{pv}, solar collection area A_{pv} [m^2], and incident solar radiation Q_t^{rad} [kW/m^2].

Solar Thermal Connector: The solar thermal collector absorbs solar radiation to heat water in hot water tank. The injected heat energy can be modeled as

$$g_t^{\text{solar}} = \eta^{\text{stc}} \cdot A_{\text{stc}} \cdot Q_t^{\text{rad}}, \quad \forall t \in T. \tag{16}$$

where η^{stc} denotes the thermal conversion efficiency, A_{stc} [m^2] is the collector area, and Q_t^{rad} [kW/m^2] represents the incident solar irradiance.

E. Multi-energy Supply-Demand Balance

The operation of the hydrogen-based MES should ensure the multi-energy demands of the building community at each time slot. This can be captured by the following balance equations

$$P_t^{\text{g}} + P_t^{\text{solar}} + P_t^{\text{fc}} + P_t^{\text{ess,dis}} = P_t^{\text{ess,ch}} + P_t^{\text{ely}} + Q_t^{\text{ED}}, \tag{17a}$$

$$g_t^{\text{solar}} + g_t^{\text{fc}} + g_t^{\text{tes,dis}} = g_t^{\text{tes,ch}} + g_t^{\text{ac}} + Q_t^{\text{HD}}, \tag{17b}$$

$$q_t^{\text{ac}} + q_t^{\text{ces,dis}} = q_t^{\text{ces,ch}} + Q_t^{\text{CD}}, \quad \forall t \in T. \tag{17c}$$

where P_t^{g} is net purchased electricity from the utility grid. $Q_t^{\text{ED}}, Q_t^{\text{HD}}$, Q_t^{CD} are electricity, heating, and cooling demand of building community.

F. Objective function

The objective is to minimize the external energy cost of the hydrogen-based MES over the optimization horizon, which can modeled as

$$J_{\text{cost}} = \mathbb{E}\left\{ \sum_{t \in T} \left[\frac{\lambda_t^{\text{b}} - \lambda_t^{\text{s}}}{2} |P_t^{\text{g}}| + \frac{\lambda_t^{\text{b}} + \lambda_t^{\text{s}}}{2} P_t^{\text{g}} + \lambda_t^{\text{h}} v_t^{\text{buy}} \right] \Delta t \right\} \tag{18}$$

where λ_t^{b} and λ_t^{s} are buying and selling price from the utility grid. λ_t^h is hydrogen trading price (buying and selling). The first two terms are electricity cost, and the third one is hydrogen cost. $\mathbb{E}$ denotes expectation for capturing the multiple uncertainties.

Overall, the optimal operation of the hydrogen-based MES can be formulated as a stochastic optimization problem $\min J_{\text{cost}}$ s.t. $(1)-(17)$, subject to nonlinear coupled system dynamics, various operational constraints and multiple sources of uncertainties. This makes model-based optimization methods computationally challenging.

III. Self-predictive Deep Reinforcement Learning

To address the computational challenge, this paper develops a deep reinforcement learning (DRL) based approach. Specifically, historical data of market prices, renewable generation, and multiple energy demands together with the proposed mathematical model are used to train an DRL agent, which is then deployed for real-time operation. It is worthy noting that the mathematical model is only used as a simulator to acquire interactive operational experience. However, the method does not depend on an explicit mathematical model if a practical system or a physics-based simulator is allowed to interact.

A. Markov Decision Process (MDP)

To develop a DRL-based approach, the problem is first required to be formulated as a Markov Decision Process (MDP). An MDP is defined by a 5-tuple (S, A, P, r, γ), where S, A, P, r, and γ denote the state space, action space, state transition probability, reward function, and discount factor, respectively. Based on the proposed mathematical model of the hydrogen-based MES, we give the following MDP formulation.

State: For the hydrogen-based MES, we define the system state at each time slot t as

$$s_t = \begin{pmatrix} day, hour, \lambda_t^{\text{b}}, \lambda_t^{\text{s}}, \lambda_t^{\text{h}}, Q_t^{\text{rad}}, \\ S_t^{\text{ess}}, S_t^{\text{tes}}, S_t^{\text{ces}}, S_t^{\text{hss}}, T_t^{\text{ely}}, C_t^{\text{ely}}, T_t^{\text{tank}}, \\ Q_t^{\text{ED}}, Q_t^{\text{HD}}, Q_t^{\text{CD}} \end{pmatrix} \tag{19}$$

where $day \in \{1, 2, \cdots, 7\}$ denotes the day of the weak and $hour \in \{1, 2, \cdots, 24\}$ denotes the hour. $\lambda_t^{\text{b}}, \lambda_t^{\text{s}}, \lambda_t^{\text{h}}$ are electricity and hydrogen market price. Q_t^{rad} is incident solar radiation for PV panels and solar thermal collectors. S_t^{ess}, S_t^{tes}, S_t^{ces}, and S_t^{hss} denote the energy levels of electrical, thermal, cooling, and hydrogen storage, respectively. $T_t^{\text{ely}}, C_t^{\text{ely}}, T_t^{\text{tank}}$ are state variables of HESS. $Q_t^{\text{ED}}, Q_t^{\text{HD}}, Q_t^{\text{CD}}$ are multi-energy demand of building community. Some studies incorporate historical information, e.g., $\lambda_{t-k:t}^{\text{b}}, \lambda_{t-k:t}^{\text{s}}, \lambda_{t-k:t}^{\text{h}}, Q_{t-k:t}^{\text{ED}}, Q_{t-k:t}^{\text{HD}}, Q_{t-k:t}^{\text{CD}}$, to enhance learning performance. This is avoided as we do not observe obvious performance gains.

Action: The action $a_t \in A$ specifies the operating policies of energy devices, which is defined as

$$a_t = \left(a_t^{\text{ess}}, a_t^{\text{tes}}, a_t^{\text{css}}, a_t^{\text{hss}}, a_t^{\text{buy}} \right) \tag{20}$$

where $a_t^{\text{ess}}, a_t^{\text{tes}}, a_t^{\text{css}}, a_t^{\text{hss}} \in [-1, 1]$ denote the normalized charging and discharging operation of electrical storage, hot water, chilled water and hydrogen tanks. $a_t^{\text{buy}} \in [-1, 1]$ indicates normalized hydrogen transaction in hydrogen market. It is noted that not all device-level control variables are explicitly included in the action as the others can be directly derived through energy balance equations given the action (20).

While interacting or in real-time operation, the normalized action should be converted into device-level operations. For the energy storage devices, the actual charging and discharging, considering physical limits can be determined by

$$\begin{aligned} P_t^{\text{ess,ch}} &= \text{clip}\Big(a_t^{\text{ess}} P_{\max}^{\text{ess,ch}}, \quad 0, \quad \frac{S_{\max}^{\text{ess}} - S_t^{\text{ess}}}{\eta^{\text{ess,ch}} \Delta t} \Big), \\ P_t^{\text{ess,dis}} &= \text{clip}\Big(-a_t^{\text{ess}} P_{\max}^{\text{ess,dis}}, 0, \frac{(S_t^{\text{ess}} - S_{\min}^{\text{ess}}) \eta^{\text{ess,dis}}}{\Delta t} \Big), \end{aligned} \tag{21}$$

$$g_t^{\text{tes,ch}} = \text{clip}\Big(a_t^{\text{tes}} g_{\max}^{\text{tes,ch}},\quad 0,\quad \frac{S_{\max}^{\text{tes}} - S_t^{\text{tes}}}{\eta^{\text{tes,ch}}\Delta t}\Big),$$
$$g_t^{\text{tes,dis}} = \text{clip}\Big(-a_t^{\text{tes}} P_{\max}^{\text{tes,dis}}, 0, \frac{(S_t^{\text{tes}} - S_{\min}^{\text{tes}})\eta^{\text{tes,dis}}}{\Delta t}\Big), \tag{22}$$
$$q_t^{\text{ces,ch}} = \text{clip}\Big(a_t^{\text{ces}} q_{\max}^{\text{ces,ch}},\quad 0,\quad \frac{S_{\max}^{\text{ces}} - S_t^{\text{ces}}}{\eta^{\text{ces,ch}}\Delta t}\Big),$$
$$q_t^{\text{ces,dis}} = \text{clip}\Big(-a_t^{\text{ces}} q_{\max}^{\text{ces,dis}}, 0, \frac{(S_t^{\text{ces}} - S_{\min}^{\text{ces}})\eta^{\text{ces,dis}}}{\Delta t}\Big), \tag{23}$$
$$v_t^{\text{hss,ch}} = \text{clip}\Big(a_t^{\text{hss}} v_{\max}^{\text{hss,ch}}, 0, \frac{S_{\max}^{\text{hss}} - S_t^{\text{hss}} - v_t^{\text{buy}}}{\Delta t}\Big),$$
$$v_t^{\text{hss,dis}} = \text{clip}\Big(-a_t^{\text{hss}} v_{\max}^{\text{hss,dis}},\ 0, \frac{S_t^{\text{hss}} + v_t^{\text{buy}} - S_{\min}^{\text{hss}}}{\Delta t}\Big), \tag{24}$$

where $\text{clip}(x,y,z) \triangleq \min\big(\max(x,y),z\big)$ bounds x within $[y,z]$. $v_{\max}^{\text{hss,dis}}, v_{\max}^{\text{hss,ch}}$ denote the maximum charging and discharging rate of the hydrogen tank. The clip operations ensure storage devices not overcharged or overdischarged during operation.

The net purchased hydrogen in hydrogen market is

$$v_t^{\text{buy}} = a_t^{\text{buy}} v_{\max}^{\text{buy}} \tag{25}$$

The operation of other energy devices can be derived from the system balance equations. Specifically, the generated cooling power and consumed heating power of AC are

$$q_t^{\text{ac}} = q_t^{\text{ces,ch}} + q_t^{\text{ces,dis}} + Q_t^{\text{CD}},$$
$$g_t^{\text{ac}} = q_t^{\text{ac}} / \eta^{\text{ac}}, \tag{26}$$

Following this, the operation of fuel cell can be sequentially inferred through the coupled thermodynamic and electrochemical dynamic equations:

$$g_t^{\text{ac}} \xrightarrow{(17b)} g_t^{\text{fc}} \xrightarrow{(13)} P_t^{\text{fc}} \xrightarrow{(12b)} i_t^{\text{fc}} \xrightarrow{(12a)} v_t^{\text{fc}} \tag{27}$$

Given the purchased hydrogen v_t^{buy}, the consumed hydrogen by fuel cell v_t^{fc}, the charging and discharging of hydrogen tank $v_t^{\text{hss,ch}}, v_t^{\text{hss,dis}}$, the produced hydrogen by electrolyzer v_t^{ely} can be derived from

$$v_t^{\text{hss,ch}} - v_t^{\text{hss,dis}} = \alpha \cdot v_t^{\text{ely}} - v_t^{\text{fc}} + v_t^{\text{buy}}, \tag{28}$$

The operating status of electrolyzer is further determined as

$$v_t^{\text{ely}} \xrightarrow{(4)} P_t^{\text{ely}} \xrightarrow{(6)} T_t^{\text{ely}} \xrightarrow{(8)} P_{t,1}^{\text{ely}} \xrightarrow{(9)} i_t^{\text{ely}} \xrightarrow{(10a)} C_t^{\text{ely}} \tag{29}$$

Finally, the amount of purchased electricity P_t^{g} can be derived from the electricity balance equation (17a).

State transition: State transitions describe the evolution of system states given the action. The transitions of time components: *day* and *hour* are deterministic and follow the calendar. The transitions of external states, including market prices $\lambda_t^{\text{b}}, \lambda_t^{\text{s}}, \lambda_t^{\text{h}}$, solar radiation Q_t^{rad} and energy demands $Q_t^{\text{ED}}, Q_t^{\text{HD}}, Q_t^{\text{CD}}$ can be captured by historical data. The storage states $S_t^{\text{ess}}, S_t^{\text{tes}}, S_t^{\text{ces}}, S_t^{\text{hss}}$, stack temperature T_t^{ely}, overload counter C_t^{ely} and tank temperature T_t^{tank} follow the system dynamics described in Section II-B and II-C.

Reward: The operation of hydrogen-based MES is to minimize the external energy cost. We define the reward as negative energy cost at each time slot. Additionally, we incorporate the operating limits of electrolyzer, fuel cell and hydrogen tank that are hard to be enforced explicitly as penalty. This leads to the following reward function

$$r_t = -\left[\frac{\lambda_t^{\text{b}} - \lambda_t^{\text{s}}}{2}|P_t^{\text{g}}| + \frac{\lambda_t^{\text{b}} + \lambda_t^{\text{s}}}{2} P_t^{\text{g}} + \lambda_t^{\text{h}} v_t^{\text{buy}}\right]\Delta t - \lambda \cdot penalty$$
$$\begin{aligned} penalty = &\left[C_t^{\text{ely}} - C_{\max}^{\text{ely}}\right]_+ + \left[T_t^{\text{ely}} - T_{\max}^{\text{ely}}\right]_+ \\ &+ \left[p_{\min}^{\text{tank}} - p_t^{\text{tank}}\right]_+ + \left[p_t^{\text{tank}} - p_{\max}^{\text{tank}}\right]_+ \\ &+ \left[P_{\min}^{\text{fc}} - P_t^{\text{fc}}\right]_+ + \left[P_t^{\text{fc}} - P_{\max}^{\text{fc}}\right]_+ \\ &+ \left[P_{\min}^{\text{ely}} - P_t^{\text{ely}}\right]_+ + \left[P_t^{\text{ely}} - P_{\max}^{\text{ely}}\right]_+ + \left[g_t^{\text{ac}} - g_{\max}^{\text{ac}}\right]_+ \end{aligned}$$

where $\lambda \geq 0$ is penalty factor for regulating the satisfaction of constraints, which is often determined by experience.

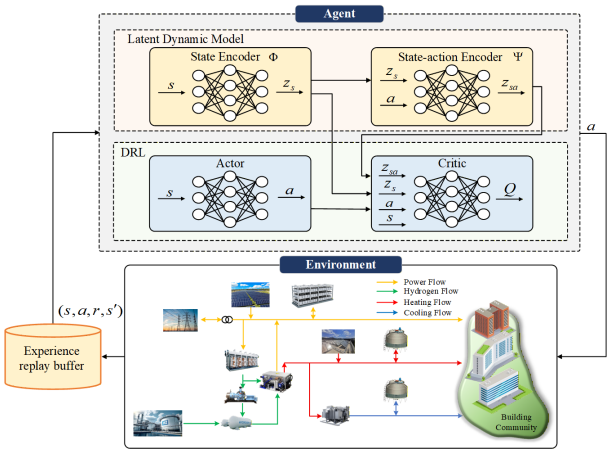

Fig. 2: The architecture of self-predictive DRL

B. Self-predictive DRL

To enable learning-based optimal operation of hydrogen-based MES, this section proposes a self-predictive deep reinforcement learning (SP-DRL) approach. The key idea is to explicitly learn a latent dynamic model of the system and leverage it to enhance the DRL learning process. Specifically, an auxiliary learning task is introduced into the conventional DRL pipeline to capture the system transition dynamics in a latent space. The learned dynamic model is then integrated into the critic network to predict the next system states, effectively guiding the action-value function learning.

This idea is directly motivated by the fact that action-value function represented by the critic network is inherently dependent on the system transition dynamics. Conventional DRL approaches typically rely on the critic network to implicitly capture the system dynamics from operational data. For complex systems or environments, the methods often suffer from the slow and unstable critic learning due to the bootstrapping effects. Since the performance of DRL is often largely determined by the critic learning, this often undermines the effectiveness of conventional DRL approaches. The proposed SP-DRL addresses this issue by explicitly learning a dynamic model of the system and incorporating it into the critic network to effectively guide critic learning in complex environments.

The implementation of SP-DRL includes two main procedures. One is to learn a latent dynamic model and the other

is to incorporate it into the DRL framework. The overall architecture of the SP-DRL is presented in Fig. 2.

Learn a latent dynamic model: A latent dynamic model can be viewed as a representation of original system dynamic model in a latent space. The benefit of learning the dynamic model in a latent space instead of original one is that it can automatically filer out task-irrelevant distractions or noises of original state space [33]. This paper employs a state-encoder and a state-action encoder (Φ,Ψ) to capture the underlying state transition dynamics in a latent space. The `state encoder` maps a raw system state s into a latent state z_s, which is described as $\Phi: S \to Z_s$. The `state-action encoder` maps the latent state and the action pair (z_s,a) into another latent state z_{sa}, which can be described as $\Psi: (S,Z_s) \to Z_{sa}$. Collectively, they form a compact latent dynamic modeling framework:

$$\begin{aligned} z_s &= \Phi(s) \\ z_{sa} &= \Psi(z_s,a) \end{aligned} \tag{30}$$

The output latent state z_{sa} can be interpreted as the predicted next latent state . Let s' denote the next state, the actual next latent state is $z_{s'}=\Phi(s')$. To ensure that (Φ,Ψ) jointly constitute a valid latent dynamic model, the predicted next latent state should be consistent with the actual next latent state, i.e.,

$$z_{sa} = z_{s'}, \forall s \in S, a \in A. \tag{31}$$

Therefore, to obtain a latent dynamic model, we define the training objective as

$$\begin{aligned} L(\Phi,\Psi) &= \mathbb{E}_{s'\sim P(\cdot|s,a)} \left\| z_{sa} - z_{s'} \right\| \\ &= \mathbb{E}_{s'\sim P(\cdot|s,a)} \left\| \Psi(\Phi(s),a) - \Phi(s') \right\| \end{aligned} \tag{32}$$

Incorporate the latent dynamic model in DRL framework: The latent dynamic model is able to predict the next system state given the current state and action, which is informative for critic learning. We therefore incorporate it into the critic network. Compared with conventional DRL, we make the following modifications of the critic network with the proposed SP-DRL method.

$$\begin{cases} DRL & \to & SP-DRL \\ \pi(s) & \to & \pi(z_s) \\ Q(s,a) & \to & Q(s,a,z_s,z_{sa}) \end{cases} \tag{33}$$

where the current latent state z_s and the predicted next latent state z_{sa} can be derived from the learned latent dynamic model (30). The proposed SP-DRL is a general framework and can be combined with any existing DRL methods that admit an actor-critic architecture, such as Deep Deterministic Policy Gradient (DDPG) [34] and Twin Delayed Deep Deterministic Policy Gradient (TD3) [35].

In SR-DRL, latent dynamic model learning and the actor-critic training of DRL are decoupled and alternated. The gradients of latent dynamic model can be computed as

$$\begin{aligned} \nabla_\Phi L(\Phi,\Psi) &= \mathbb{E}_{s'\sim P(\cdot|s,a)} \left[\nabla_\Phi \left\| \Psi(\Phi(s),a) - \left|\Phi(s')\right|_\times \right\| \right] \\ \nabla_\Psi L(\Phi,\Psi) &= \mathbb{E}_{s'\sim P(\cdot|s,a)} \left[\nabla_\Psi \left\| \Psi(\Phi(s),a) - \left|\Phi(s')\right|_\times \right\| \right] \end{aligned} \tag{34}$$

Algorithm 1 Training procedures of self-predictive DRL

Initialization: (1) Initialize actor network π^θ, $\pi^{\theta'}$, critic Q^ω, $Q^{\omega'}$, state encoder Φ, state-action encoder Ψ; (2) Initialize replay buffer D.
for episode $= 1$ to N_{episode} **do**
  Reset environment and obtain initial state s_0; initialize random noise $\mathcal{N}_t$
  for time step $t = 0$ to $T-1$ **do**
    Select action: $a_t = \pi^\theta(s_t) + \mathcal{N}_t$
    Execute a_t and get r_t, s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in D
    if $|D| > D_{\min}$ **then**
      Sample a mini-batch $(s_t, a_t, r_t, s_{t+1}) \sim D$
      Update encoders according to (34)
      Update critic according to (35) or (36)
      Update actor according to (37)
      Soft update target networks θ', ω'
    end if
  end for
end for

where $|\cdot|_\times$ denotes stop-gradient operation, which is used to avoid the collapse of representation learning during training.

For actor and critic updates, we denote the actor and target actor networks as $\pi^\theta, \pi^{\theta'}$, the critic and target critic networks as $Q^\omega, Q^{\omega'}$. By adopting the idea of conventional DRP, the critic is trained to minimize the Bellman error and the actor is updated to maximize the action-value function. When combined with DDPG and TD3, the updates of critic and actor networks can be described as

a) DDPG Critic Updates:

$$\begin{cases} L_{\text{critic}}(\omega) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}\left[\left(Q^\omega(s_t,a_t,z_s,z_{sa}) - y_t \right)^2 \right], \\ y_t = r_t + \gamma Q^{\omega'}\left(s_{t+1}, \pi^{\theta'}(s_{t+1}), z_{s'}, z_{s'a'}\right), \\ \omega \leftarrow \omega - \eta_Q \nabla_\omega L_{\text{critic}}(\omega). \end{cases} \tag{35}$$

b) TD3 Critic Updates:

$$\begin{cases} L_{\text{critic}}(\omega_i) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}\left[\left(Q^{\omega_i}(s_t,a_t,z_s,z_{sa}) - y_t \right)^2 \right], \\ y_t = r_t + \gamma \min_{i=1,2} Q^{\omega_i'}\left(s_{t+1}, \pi^{\theta'}(s_{t+1}) + \varepsilon, z_{s'}, z_{s'a'}\right), \\ \omega_i \leftarrow \omega_i - \eta_Q \nabla_{\omega_i} L_{\text{critic}}(\omega_i), \quad i = 1,2. \end{cases} \tag{36}$$

where $\varepsilon \sim \text{clip}(\mathcal{N}(0,\sigma), -c, c)$ is the clipped policy noise, and $Q^{\omega_1'}, Q^{\omega_2'}$ are double target critic networks with TD3.

c) DDPG/TD3 Actor Updates::

$$\begin{cases} \nabla_\theta J \approx \mathbb{E}_{s\sim D}\left[\nabla_a Q^\omega(s_t,a_t,z_s,z_{sa})\big|_{a=\pi^\theta(s)} \nabla_\theta \pi^\theta(s) \right], \\ \theta \leftarrow \theta + \eta_\pi \nabla_\theta J. \end{cases} \tag{37}$$

where D represents the experience replay buffer. η^Q and η^π are critic and actor update stepsize.

The target actor and critic networks of DRL are often updated at a much slow space for stable training. The widely used soft update scheme is adopted in our framework. The detailed implementation of the SR-DRL is summarized in Algorithm 1. After finishing training, the trained DRL agents can be deployed for the real-time operation of the hydrogen-based MES. At each operating point, the operation of energy

devices can be obtained by the trained DRL agent with the observed system state.

IV. Case Studies

This section evaluates the proposed SP-DRL methods for hydrogen-based MES. We first examine the convergence of latent dynamic model learning. Next, we investigate the training performance of the SP-DRL agents. Finally, we assess the operational performance of the trained agents deployed for real-time operation of the hydrogen-based MES.

A. Simulation Setup

Real-world data is used to set up case studies. Multi-energy demand profiles (electricity, heating, and cooling) of building community, along with solar radiation, weather conditions and the time-of-use (ToU) utility price are obtained from the CityLearn dataset [36]. A hydrogen system comprising a Siemens Silyzer 100 electrolyzer, a Swiss hydrogen fuel cell, a compressor, and a pressurized hydrogen storage tank of [24, 37] is adopted, with the configurations slightly adapted to accommodate the multi-energy demands of the community. The settings of other energy generation, conversion and storage devices follow [10, 38]. To simplify discussions, hydrogen market is omitted in our case studies. The penalty factors of constraints are set to 1.0 by experience. We refer the readers to our extended version [39] for the detailed system configurations. We compare the proposed SP-DRL methods against their conventional counterparts as follows.

- `DDPG` and `SP-DDPG`: Conventional DDPG and the proposed self-predictive DDPG are trained on real-world data for the hydrogen-based MES.
- `TD3` and `SP-TD3`: Conventional TD3 and the proposed self-predictive TD3 are trained on real-world data for the hydrogen-based MES.

Three cases (`Case 1`, `Case 2`, and `Case 3`), corresponding to different data periods, are considered. For each case, the DRL agents are first trained on 20-day dataset and subsequently tested on another 10-day dataset. For a fair comparison, all DRL methods share identical model architectures and training hyperparameters, except for the state and state-action encoders with the SP-DRL methods. We refer the readers to our extended version [39] for detailed model architecture and training hyperparameters.

B. Latent Dynamic Model Learning

This sections evaluates the convergence of latent dynamic model learning together with its predictive performance. In terms of convergence, the training loss (32) of state and state-action encoders during the training process of `Case 1` is first investigated. As shown in Fig. 3 (a), the training loss decreases quickly towards zero, indicating the convergence of latent dynamic model learning. After finishing the training, we further examine the learned latent dynamic model for predicting system states in latent space. Specifically, we use the trained state and state-action encoders to predict the next latent system states and compare them with the actual ones on the 20-day training data. For each time step, we compute the actual next latent state by $z_{s'} = \Phi(s')$ and the predicted next latent state by $z_{sa} = \Psi(\Phi(s), a)$. To handle the high-dimensional latent system states, we randomly generate some directional vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ and use them to project the high-dimensional latent states into one-dimensional ones by a multiplication operation. Fig. 3 (b)–(d) presents the projections of actual next latent states $z_{s'}$ and the predicted ones z_{sa} across the time. We observe that the two trajectories under the randomly generated projection vectors coincide with each other across the time. This indicates that the learned latent dynamic model can well predict the next system state in latent space given the current state and action. Similar convergence and predictive performance of latent dynamic model are observed in `Case 2` and `Case 3`. This demonstrates that the proposed state and state-action encoder can effectively learn a dynamic model for capturing the underlying system dynamics of the hydrogen-based MES.

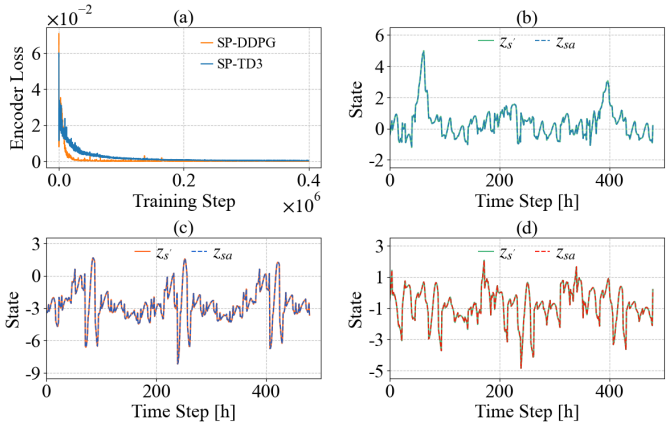

Fig. 3: Convergence of latent dynamics learning and its predictive capability.

C. Training Performance

This section investigates the training performance of the proposed SP-DRL methods in comparison with their conventional counterparts. First of all, the episode returns of the DRL methods during the training process across the three cases are investigated. As shown in Fig. 4, we observe that the SP-DRL methods (`SP-TD3` and `SP-DDPG`) consistently outperform their conventional counterparts (`TD3` and `DDPG`). Specifically, `SP-TD3` and `SP-DDPG` achieve faster convergence rate and significantly higher final episode returns compared with the conventional methods. Moreover, the SP-DRL methods exhibit much more stable training behavior as reflected by the smooth episode return trajectories.

The enhanced training performance of SP-DRL methods can be attributed to the improved critic learning enabled by the learned latent dynamic model. This can be perceived from Fig. 5, which shows the normalized TD error of critic networks during the training process for `Case 1`. It is observed that the TD error decreases much faster towards zero with the `SP-TD3` and `SP-DDPG` compared to their conventional counterparts `TD3` and `DDPG`. Beyond that, we observe obvious fluctuations of TD errors with the `TD3` and `DDPG` during the training process. This is actually the cause of the unstable episode returns with `TD3` and `DDPG` observed in Fig. 4. These results demonstrate that by learning a latent dynamic model to

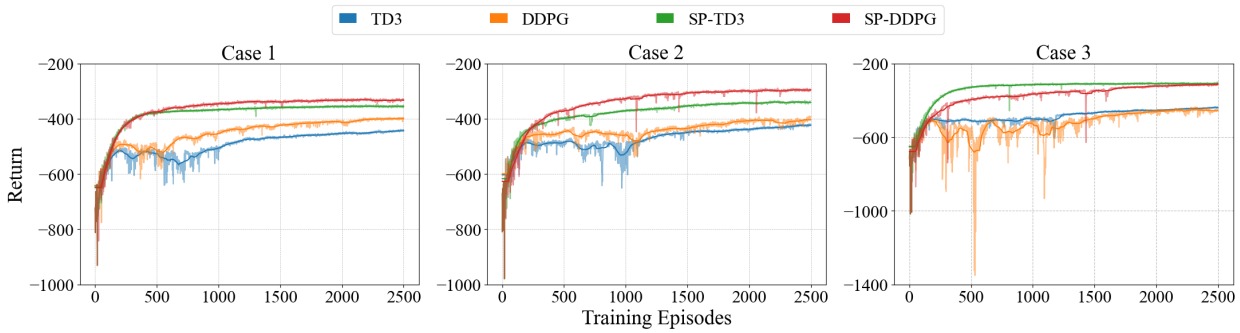

Fig. 4: The evolution of episode return during training under different methods.

explicitly predict the next system state for the critic network can effectively enhance the critic learning, thereby enhancing the performance of the DRL methods.

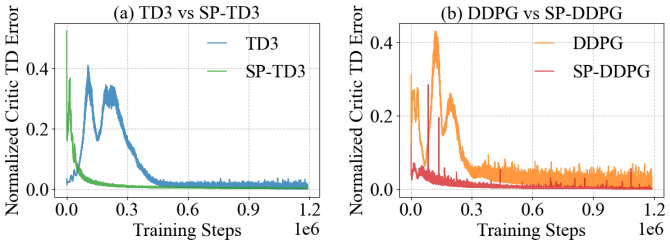

Fig. 5: Normalized TD error of critic networks during the training process under different DRL methods in `Case 1`.

TABLE I: Performance of different DRL agents on training and testing datasets

Cases	Methods	Training dataset		Testing dataset	
		Avg. Cost [$]	Perf. Imp.[%]	Avg. Cost [$]	Perf. Imp.[%]
Case 1	TD3	398.68	–	551.73	–
	SP-TD3	**320.16**	**19.7**	**470.05**	**14.8**
	DDPG	351.54	–	529.49	–
	SP-DDPG	**304.91**	**13.3**	**456.41**	**13.8**
Case 2	TD3	372.38	–	292.39	–
	SP-TD3	**302.26**	**18.8**	**236.95**	**19.0**
	DDPG	315.92	–	253.83	–
	SP-DDPG	**266.12**	**15.8**	**225.66**	**11.1**
Case 3	TD3	405.29	–	248.08	–
	SP-TD3	**276.36**	**31.8**	**185.25**	**25.3**
	DDPG	345.38	–	230.97	–
	SP-DDPG	**287.34**	**16.8**	**190.86**	**17.4**

D. Operational Performance

This section evaluates the operational performance of trained DRL agents deployed on the hydrogen-based MES. For each of the trained DRL agents, we examine their induced average daily energy cost on both the 20-day training dataset and the 10-day testing dataset across the three cases and use them as performance metrics. For each case, we evaluate the performance improvement of the SP-DRL methods relative to their conventional counterparts. TABLE I reports the obtained results. For ease of comparison, we have highlighted the the results of SP-DRL methods in bold. From the results, we observe that the SP-DRL methods (`SP-TD3` and `SP-DDPG`) consistently provide considerably lower average daily energy cost compared with their conventional counterparts (`TD3` and `DDPG`) across all tested cases. Specifically, for `Case 1`, the average daily energy cost of hydrogen-based MES is reduced by 19.7% (`SP-TD3`) and 13.3% (`SP-DDPG`) on the training dataset, and by 14.8% (`SP-TD3`) and 13.8% (`SP-DDPG`) on the testing dataset. Obvious performance improvement are observed in `Case 2` and `Case 3`. Overall, the SP-DRL methods provide about 11-32% performance improvement over conventional DRL methods across all tested cases. Besides, it is worthy noting that each of the trained DRL agents yields relatively close performance on the training and testing datasets. This indicates the strong generalization capability and stable performance of the trained DRL agents.

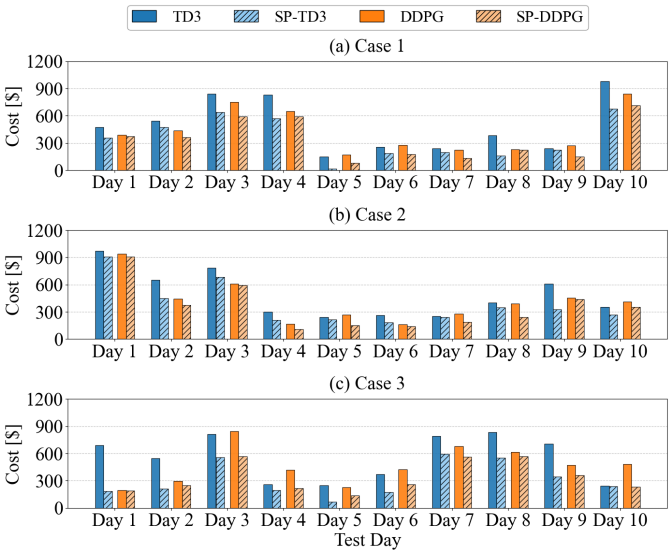

Fig. 6: Distribution of the daily energy cost of hydrogen-based MES under the different DRL agents on testing dataset.

The improved operational performance of SP-DRL methods over the conventional DRL can be further perceived from Fig. 6, which presents the distributions of daily energy cost over the 10-day testing dataset across the three cases. It can be observed that the SP-DRL methods (`SP-TD3` and `SP-DDPG`) achieve considerably lower daily energy cost compared with the conventional DRL approaches (`TD3` and `DDPG`) on all the tested days.

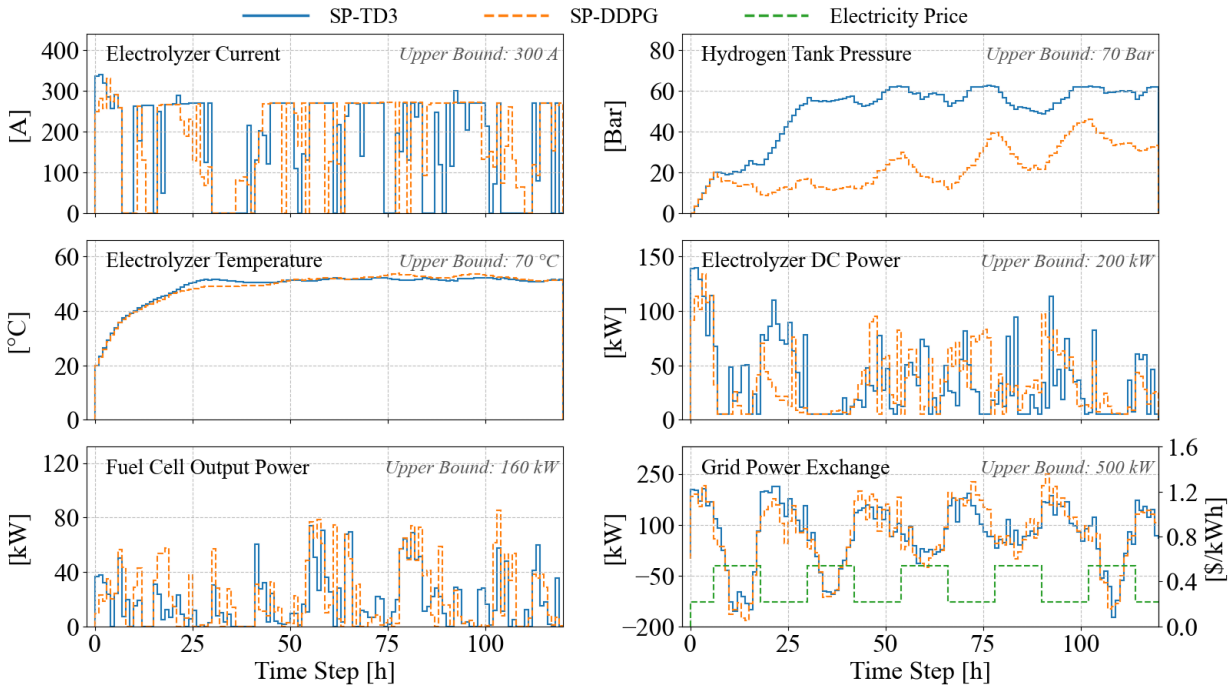

Fig. 7: Operational profiles of the devices from the hydrogen energy storage system over 5 days under `SP-TD3` and `SP-DDPG`.

E. Operational Policies

This section evaluates the effectiveness of operational policies provided by the trained SP-DRL agents. We focus on the hydrogen system, including the electrolyzer, fuel cell and hydrogen as it represents a central component of the hydrogen-based MES and exhibits complicated operating characteristics. Fig. 7 presents the operational trajectories of the hydrogen system in `Case 1` over the first 5 days of the testing dataset. Specifically, the electrolyzer current, hydrogen tank pressure, electrolyzer temperature, electrolyzer DC power, fuel cell output power and utility grid transactions under SP-DRL methods (`SP-TD3` and `SP-DDPG`) are presented. The results show that all variables are controlled within their prescribed physical limits throughout the period. Specifically, the hydrogen tank pressure stays below the upper limit of 70 bar. The electrolyzer temperature maintains close to the prescribed temperature of 60 °C. The electrolyzer injected DC power and fuel cell output power remain below their respective upper bounds of 200 kW and 160 kW. It is worth noting that the electrolyzer current temporarily exceeds its nominal current 300 A. This is because the electrolyzer is allowed to be overloaded for some period without causing any adverse effects. In addition, we observe that grid power purchases of the hydrogen-based MES mainly occur during low-price periods, and some grid injections are observed during high-price periods. This is reasonable considering the external energy cost reduction objective. The above results demonstrate that the DRL methods can provide effective operational policies for hydrogen-based MES.

V. Conclusion

This paper studied the optimal operation of a hydrogen-based multi-energy system (MES) comprising diverse energy generation, conversion, and storage devices to provide coordinated electricity, heating, and cooling energy to a building community. We proposed a comprehensive operational optimization model for the system that fully capture the nonlinear and coupled thermal, electrochemical and pressure dynamics of electrolyzers, fuel cells and hydrogen tank to enable a reliable operation of the system. We further developed a self-predictive deep reinforcement learning (SP-DRL) method to enable learning-based optimal operation of the system. The method learned a latent dynamic model to capture the underlying system dynamics to facilitate the critic learning, thereby enhancing the DRL learning. Case studies based on real-world data demonstrated that the proposed SP-DRL method can significantly improve the convergence rate, sampling efficiency, training stability and policy performance compared with conventional DRL methods. The average daily energy cost of hydrogen-based MES was reduced by about 11-32% across all tested scenarios.

References

[1] International Energy Agency, "Net zero by 2050: A roadmap for the global energy sector," 2021. Available: https://www.iea.org/reports/net-zero-by-2050.

[2] L. Shi, X. Chen, and R. Jing, "Muti-energy integration towards energy decarbonization," 2023.

[3] R. Last Name, S. Coauthor, I. Coauthor, S. Coauthor, and A. Coauthor, "A critical review of green hydrogen production by electrolysis: From technology and modeling to performance and cost," *Energies*, vol. 19, no. 1, p. 59, 2026.

[4] S. Chen, J. Zhang, Z. Wei, H. Cheng, and S. Lv, "Towards renewable-dominated energy systems: role of green hydrogen," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 6, pp. 1697–1709, 2024.

[5] M. Shaterabadi, S. Sadeghi, and M. A. Jirdehi, "The role of green hydrogen in achieving low and net-zero carbon emissions: climate change and global warming," in *Green hydrogen in power systems*, pp. 141–153, Springer, 2024.

[6] X. Yang, C. P. Nielsen, S. Song, and M. B. McElroy, "Breaking the hard-to-abate bottleneck in china's path to carbon neutrality with clean hydrogen," *Nature Energy*, vol. 7, no. 10, pp. 955–965, 2022.

[7] X. Guan, W. Guo, Z. Xu, J. Liu, and J. Wu, "Cost efficient operation of a hydrogen enabled zero-carbon multi-energy system," *IEEE Transactions on Sustainable Energy*, 2026.

[8] Y. Sun, H.-W. Li, D. Wang, and C.-H. Du, "A novel zero carbon emission system based on the complementary utilization of solar energy and hydrogen," *Applied Energy*, vol. 356, p. 122443, 2024.

[9] J. Liu, Z. Xu, J. Wu, K. Liu, and X. Guan, "Optimal planning of distributed hydrogen-based multi-energy systems," *Applied Energy*, vol. 281, p. 116107, 2021.

[10] X. Dong, J. Wu, Z. Xu, K. Liu, and X. Guan, "Optimal coordination of hydrogen-based integrated energy systems with combination of hydrogen and water storage," *Applied energy*, vol. 308, p. 118274, 2022.

[11] J. Liu, Y. Zhao, Z. Xu, J. Wu, K. Liu, and X. Guan, "Coordinated integration of hydrogen-enabled multi-energy systems to achieve zero-carbon buildings," *Building and Environment*, p. 113433, 2025.

[12] X. Fang, W. Dong, Y. Wang, and Q. Yang, "Multiple time-scale energy management strategy for a hydrogen-based multi-energy microgrid," *Applied Energy*, vol. 328, p. 120195, 2022.

[13] Z. Chen, L. Yu, M. Chen, D. Yue, T. Zhang, Y. Ye, G. Strbac, and M. Zhang, "Reliability and comfort-aware operation optimization for hydrogen-based building energy systems in off-grid mode," *IEEE Transactions on Smart Grid*, 2025.

[14] Q. Sui, H. He, J. Liang, Z. Li, and C. Su, "Short-term scheduling of integrated electric-hydrogenthermal systems considering hydroelectric power plant peaking for hydrogen vessel navigation," *IEEE Transactions on Sustainable Energy*, 2025.

[15] S. Mullanu, C. Chua, A. Molnar, and A. Yavari, "Artificial intelligence for hydrogen-enabled integrated energy systems: A systematic review," *International Journal of Hydrogen Energy*, vol. 141, pp. 283–303, 2025.

[16] A. Franzoso, G. Fambri, and M. Badami, "Deep reinforcement learning as a tool for the analysis and optimization of energy flows in multi-energy systems," *Energy Conversion and Management*, vol. 341, p. 120095, 2025.

[17] A. Dolatabadi, H. Abdeltawab, and Y. A.-R. I. Mohamed, "A novel model-free deep reinforcement learning framework for energy management of a pv integrated energy hub," *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4840–4852, 2022.

[18] L. Zhao, T. Yang, W. Li, and A. Y. Zomaya, "Deep reinforcement learning-based joint load scheduling for household multi-energy system," *Applied Energy*, vol. 324, p. 119346, 2022.

[19] L. Zhang, Y. He, H. Wu, and N. D. Hatziargyriou, "An optimal scheduling framework for integrated energy systems using deep reinforcement learning and deep learning prediction models," *IEEE Transactions on Smart Grid*, 2025.

[20] T. Zhang, M. Sun, D. Qiu, X. Zhang, G. Strbac, and C. Kang, "A bayesian deep reinforcement learning-based resilient control for multi-energy micro-gird," *IEEE Transactions on Power Systems*, vol. 38, no. 6, pp. 5057–5072, 2023.

[21] B. Zhang, W. Hu, A. M. Ghias, X. Xu, and Z. Chen, "Multi-agent deep reinforcement learning based distributed control architecture for interconnected multi-energy microgrid energy management and optimization," *Energy Conversion and Management*, vol. 277, p. 116647, 2023.

[22] H. Dong, Z. Shan, J. Zhou, C. Xu, and W. Chen, "Refined modeling and co-optimization of electric-hydrogen-thermal-gas integrated energy system with hybrid energy storage," *Applied Energy*, vol. 351, p. 121834, 2023.

[23] M. Fochesato, C. Peter, L. Morandi, and J. Lygeros, "Peak shaving with hydrogen energy storage: From stochastic control to experiments on a 4 MWh facility," *Applied Energy*, vol. 376, p. 123965, 2024.

[24] B. Flamm, C. Peter, F. N. Büchi, and J. Lygeros, "Electrolyzer modeling and real-time control for optimized production of hydrogen gas," *Applied Energy*, vol. 281, p. 116031, 2021.

[25] N. Qi, K. Huang, Z. Fan, and B. Xu, "Long-term energy management for microgrid with hybrid hydrogen-battery energy storage: A prediction-free coordinated optimization framework," *Applied Energy*, vol. 377, p. 124485, 2025.

[26] S. Fujimoto, W.-D. Chang, E. Smith, S. S. Gu, D. Precup, and D. Meger, "For sale: State-action representation learning for deep reinforcement learning," *Advances in neural information processing systems*, vol. 36, pp. 61573–61624, 2023.

[27] R. Sun, H. Zang, X. Li, and R. Islam, "Learning latent dynamic robust representations for world models," *Proceedings of Machine Learning Research*, vol. 235, pp. 47234–47260, 2024.

[28] A. Zhang, R. T. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," in *International Conference on Learning Representations*.

[29] A. X. Lee, A. Nagabandi, P. Abbeel, and S. Levine, "Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model," *Advances in neural information processing systems*, vol. 33, pp. 741–752, 2020.

[30] M. Alles, P. Becker-Ehmck, P. van der Smagt, and M. Karl, "Constrained latent action policies for model-based offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 70381–70405, 2024.

[31] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn, "Offline reinforcement learning from images with latent space models," in *Learning for dynamics and control*, pp. 1154–1168, PMLR, 2021.

[32] J. Assaf and B. Shabani, "Transient simulation modelling and energy performance of a standalone solar-hydrogen combined heat and power system integrated with solar-thermal collectors," *Applied Energy*, vol. 178, pp. 66–77, 2016.

[33] C. Voelcker, T. Kastner, I. Gilitschenski, and A.-m. Farahmand, "When does self-prediction help? understanding auxiliary tasks in reinforcement learning," *arXiv preprint arXiv:2406.17718*, 2024.

[34] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[35] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1587–1596, 2018.

[36] K. Nweye, K. Kaspar, G. Buscemi, T. Fonseca, G. Pinto, D. Ghose, S. Duddukuru, P. Pratapa, H. Li, J. Mohammadi, *et al.*, "Citylearn v2: energy-flexible, resilient, occupant-centric, and carbon-aware management of grid-interactive communities," *Journal of Building Performance Simulation*, vol. 18, no. 1, pp. 17–38, 2025.

[37] M. Fochesato, C. Peter, L. Morandi, and J. Lygeros, "Peak shaving with hydrogen energy storage: From stochastic control to experiments on a 4 mwh facility," *Applied Energy*, vol. 376, p. 123965, 2024.

[38] J. Liu, Z. Xu, J. Wu, K. Liu, and X. Guan, "Optimal planning of distributed hydrogen-based multi-energy systems," *Applied Energy*, vol. 281, p. 116107, 2021.

[39] Z. Pu, Y. Yu, *et al.*, "Optimal operation of hydrogen-based multi-energy system via a self-predictive deep reinforcement learning approach." https://yangyu-bears-berkeley.github.io/pdf/SR-DRL-Hydrogen-Energy_system.pdf, 2026. Preprint.

APPENDIX A. PARAMETERS

TABLE II: Energy system configuration and device parameters

Device	Param.	Value	Units	Param.	Value	Units
ESS	$S^{\text{ess,min}}$	0	kWh	$S^{\text{ess,max}}$	50	kWh
	$P^{\text{ess,ch}}_{\text{max}}$	20	kW	$P^{\text{ess,dis}}_{\text{max}}$	20	kW
	$S^{\text{ess}}_{\text{init}}$	5	kWh	$\eta^{\text{ess,ch/dis}}$	0.95	–
TES	$S^{\text{tes,min}}$	0	kWh	$S^{\text{tes,max}}$	100	kWh
	$S^{\text{tes}}_{\text{init}}$	10	kWh	$g^{\text{tes,ch}}_{\text{max}}$	50	kW
	$g^{\text{tes,dis}}_{\text{max}}$	50	kW	$\eta^{\text{tes,ch/dis}}$	0.9	–
CES	$S^{\text{ces,min}}$	20	kWh	$S^{\text{ces,max}}$	200	kWh
	$S^{\text{ces}}_{\text{init}}$	20	kWh	$q^{\text{ces,ch}}_{\text{max}}$	40	kW
	$q^{\text{ces,dis}}_{\text{max}}$	40	kW	$\eta^{\text{ces,ch/dis}}$	0.9	–
EL	$P^{\text{ely}}_{\text{min}}$	5	kW	$P^{\text{ely}}_{\text{max}}$	200	kW
	$P^{\text{ely}}_{\text{nom}}$	100	kW	$P^{\text{ely}}_{\text{nom,1}}$	25	kW
	z_1	$1.618 \cdot 10^{-5}$	$\text{m}^3/[^\circ\text{C} \cdot s]$	z_0	$1.490 \cdot 10^{-2}$	m^3/s
	z_{low}	$1.530 \cdot 10^{-4}$	$\text{m}^3/[\text{s} \cdot \text{kW}]$	z_{high}	$1.195 \cdot 10^{-4}$	$\text{m}^3/[\text{s} \cdot \text{kW}]$
	$T^{\text{ely}}_{\text{max}}$	70	$^\circ$C	$T^{\text{ely}}_{\text{init}}$	20	$^\circ$C
	j_0	3.958	$^\circ$C	j_1	0.551	–
	j_2	0.430	$^\circ\text{C}/\text{kWh}$	h_0	235.254	$^\circ$C
	h_1	0.673	–	h_{low}	0.987	A/kW
	h_{high}	9.0	A/kW	$i^{\text{ely}}_{\text{nom}}$	300	A
	$C^{\text{ely}}_{\text{max}}$	120	Ah	α	0.697	–
HT	$S^{\text{hss,min}}$	0	kg	$S^{\text{hss,max}}$	50	kg
	$S^{\text{hss}}_{\text{init}}$	0	kg	V^{tank}	10	m^3
	$p^{\text{tank}}_{\text{min}}$	0	bar	$p^{\text{tank}}_{\text{max}}$	70	bar
	$m^{\text{hss,ch}}_{\text{max}}$	30	m^3/h	$m^{\text{hss,dis}}_{\text{max}}$	30	m^3/h
	$\eta^{\text{hss,ch/dis}}$	1.0	–	ρ_0	$8.99 \cdot 10^{-2}$	kg/m^3
	b_0	$11.5 \cdot 10^5$	m^2/s^2	b_1	$4.16 \cdot 10^3$	$\text{m}^2/[^\circ\text{C} \cdot \text{s}^2]$
	g_0	0.94	–	g_1	$5.91 \cdot 10^{-2}$	–
FC	$P^{\text{fc}}_{\text{max}}$	160	kW	$P^{\text{fc}}_{\text{min}}$	0	kW
	$P^{\text{fc}}_{\text{bp}}$	47.97	kW	η_{fc}	0.3	–
	$\eta^{\text{fc}}_{\text{rec}}$	0.8	–	$i^{\text{fc}}_{\text{bp}}$	122.80	A
	s_1	2.56	1/kV	s_2	3.31	1/kV
	c	0.21	Nm^3/C	–	–	–
AC	$g^{\text{ac}}_{\text{max}}$	200	kW	η_{ac}	0.94	–
PV	A_{PV}	1500	m^2	A_{stc}	400	m^2
	η^{stc}	0.762	–	η^{PV}	0.2	–

TABLE III: DRL and SR-DRL model and training parameters

Parameter	Value	Parameter	Value
Optimizer	Adam	Epochs	2500
Batch size	1024	Buffer size	10^6
Discount rate	1.0	Learning rate	3×10^{-4}
Soft update rate	0.01		
State embedding	[256, 256, 512]	State-action embedding	[256, 256, 512]
Actor	[512, 512, 512, action_dim]	Critic	[512, 512, 1]