ASME Journal of Manufacturing Science and Engineering
Online journal at:
https://asmedigitalcollection.asme.org/manufacturingscience

Muhammad Waseem
Department of Mechanical and Aerospace Engineering,
University of Virginia,
Charlottesville, VA 22903
e-mail: kqr5pu@virginia.edu

Mihitha Sarinda Maithripala
Department of Electrical and Computer Engineering,
University of Virginia,
Charlottesville, VA 22903
e-mail: wpg8hm@virginia.edu

Qing Chang[1]
Senior Mem. ASME
Professor
Department of Mechanical and Aerospace Engineering,
University of Virginia,
Charlottesville, VA 22903
e-mail: qc9nq@virginia.edu

Zongli Lin
Senior Mem. ASME
Professor
Department of Electrical and Computer Engineering,
University of Virginia,
Charlottesville, VA 22903
e-mail: zl5y@virginia.edu

Integrated Energy Optimization in Manufacturing Through Multiagent Deep Reinforcement Learning: Holistic Control of Manufacturing, Microgrid Systems, and Battery Storage

Microgrid technology integrates storage devices, renewable energy sources, and controllable loads and has been widely explored in residential, commercial, and critical facilities. However, its potential in manufacturing remains largely underexplored, where optimal control of microgrids containing energy storage systems (ESS) is crucial. Two primary challenges arise in integrated microgrid-manufacturing systems: fluctuating renewable energy output and nondeterministic polynomial (NP)-hard demand-side control. Addressing both challenges simultaneously increases complexity. This article proposes an integrated control considering ESS degradation, optimizing control on both the manufacturing demand and microgrid energy supply sides within the production constraints. It formulates the problem in a decentralized partially observable Markov decision process (Dec-POMDP) framework, treating the system as a multiagent environment. The multiagent deep deterministic policy gradient (MADDPG) algorithm is adapted to optimize control policies. Investigating the trained policies provides insights into their logic, and a rule-based policy is introduced for practical implementation. Experimental validation on a manufacturing system validates the effectiveness of the proposed method and the rule-based policy. [DOI: 10.1115/1.4067614]

Keywords: microgrid, battery degradation, manufacturing, scheduling, deep learning, multiagent, control and automation, modeling and simulation, sustainable manufacturing

1 Introduction

A microgrid is a group of interconnected loads and distributed energy resources that act as a single controllable entity with respect to the grid. The concept of "microgrid" remains fluid and ever-evolving, offering both challenges and opportunities [1]. Recent advancements in software systems and reductions in energy technology costs, along with consumer demands for sustainability and reliability, have propelled microgrids into a new era of capability and deployment. These interconnected power systems feature autonomous control and can operate both independently and in conjunction with the main power grid, offering reliability, cost-effectiveness, and emission reduction [2]. This surge in academic discourse surrounding microgrid implementation underscores the urgent need for environmental preservation.

Typically, a central controller oversees energy management in microgrids, allocating solar, wind, and generator energy based on the overall demand. Battery technology advancements have increased capacity, allowing surplus energy to serve as flexible resources [3]. This diversification offers opportunities for optimizing energy allocation but poses challenges such as degradation of energy storage systems (ESS), synergy among energy sources, and managing supply diversity. The longevity of an ESS is influenced by various factors, such as the depth of discharge, discharge rate, duration at low and high states of charge, current fluctuations, and the extent and frequency of overcharging. While several studies have addressed these issues independently [4,5], these factors are often ignored in an integrated system where besides the demand and supply, the ESS should be handled in an optimum way, which not only improves the ESS lifetime but also plays a significant role toward sustainability.

The advantages of an integrated demand-supply-storage system can be examined from several angles. First, microgrids provide an additional energy supply option, enhancing cost-effectiveness by offering flexibility in energy source selection and flow regulation to optimize efficiency. Moreover, microgrid utilization strengthens resilience against utility failures, such as those caused by natural disasters. Second, controlling manufacturing on the demand side

[1]Corresponding author.

Manuscript received June 25, 2024; final manuscript received December 28, 2024; published online February 11, 2025. Assoc. Editor: Ran Jin.

reduces operational costs and enhances production throughput, leading to lower unit costs. Third, integrating microgrids and ESS contributes to environmental sustainability by incorporating renewable energy sources (RES) and optimizing ESS management, significantly reducing the carbon footprint of supplied energy.

Several studies have explored integrating manufacturing and microgrid systems, employing reinforcement learning (RL) algorithms for joint control of energy distribution and machine operations [6,7]. However, optimizing control in complex systems with multiple inputs necessitates improved coordination among them. The diverse energy sources in microgrids add complexity, requiring coordination due to multiple supply directions. To consider coordination, Ref. [7] addressed a similar integrated system using a multiagent approach. However, their focus was solely on controlling the microgrid side, without considering the dynamic demand, specifically the manufacturing operation aspect highlighted in this article and the storage system. Effective utilization of multiagent reinforcement learning (MARL)-based control in such situations requires a clearly defined and formulated control problem within a suitable framework.

Therefore, this article introduces a unified control framework employing a decentralized partially observable Markov decision process (Dec-POMDP) that considers the supply (microgrid), storage (ESS), and demand (manufacturing system) dynamics. A multiagent deep deterministic policy gradient (MADDPG) algorithm is adopted to address both the discrete and continuous optimal control actions. Experiments are conducted on a manufacturing system equipped with ESS and an onsite microgrid with renewable sources, utilizing real parameters to determine optimal control actions for both the manufacturing system and microgrid to achieve cost optimization. Empirical results validate that the optimal policies derived from the proposed control algorithm enhance production efficiency and reduce costs compared to randomly sampled policies and standard operational procedures.

In summary, the main contributions of the proposed joint control for integrated microgrid-manufacturing systems are as follows:

- Modeling and evaluating the dynamics of combined energy consumption, covering microgrids, ESS, and manufacturing systems.
- Framing the supervisory control of the integrated microgrid-manufacturing system within the Dec-POMDP framework as a multiagent system comprising both discrete and continuous agents.
- Employing the MADDPG deep reinforcement learning (DRL) algorithm to tackle the control challenge. Following training, the policy is assessed, and a rule-based approach is extracted for convenient implementation in real-world scenarios, sidestepping the requirement for intricate computational resources.

The rest of the article is structured as follows. In Sec. 2, the literature review is conducted. In Sec. 3, the system description is provided. The system modeling is covered in Sec. 4, followed by the problem formulation in Sec. 5. The results of the case study, which compares the proposed method to other control methods, are presented in Sec. 6. Section 7 introduces and discusses the proposed rule-based control. Finally, the conclusion is given in Sec. 8.

2 Literature Review

Microgrids connect various independent and diverse renewable energy systems to create a complex and dynamic integrated energy system, essentially a system of systems. The basic structure of a microgrid includes generators (renewable or nonrenewable), storage systems, and loads. It can operate in alternating current, direct current, or a combination of both, each with its advantages and disadvantages [8]. Despite the dynamic nature of renewable energy resources, with appropriate balance and control, a complex renewable energy microgrid can provide stable and satisfactory electricity and energy.

Apart from the efforts on the energy supply side, efficient energy usage is also imperative on the demand side. While residential microgrid literature [9] extensively covers interconnection structures among different units, there is limited focus on the manufacturing sector [6,7]. The industrial sector stands as the largest global energy consumer, accounting for 52% of total energy consumption [10]. In manufacturing, uninterrupted power is vital for operations, leading to studies aiming to optimize microgrid designs and component sizing. For example, one study proposed a model for sizing onsite generation systems considering energy demands from manufacturing and heating, ventilation, and air conditioning (HVAC) systems [11]. Another study forecasted energy load fluctuations and solar availability, evaluating onsite microgrid costs [12]. However, existing studies often oversimplify manufacturing systems, neglecting system dynamics.

Considering the complex dynamics, particularly the uncertainties of renewable sources and load involved in microgrid operation, model predictive control (MPC)-based approaches have been widely used to estimate uncertainty and arrive at an optimizer to solve the best schedules of microgrid operations [13]. For example, an online optimal energy management model for energy storage systems in a microgrid was developed using mixed integer linear programming over a rolling horizon period [14]. A similar method considering the time-varying constraints of a microgrid was proposed in Ref. [15]. In Ref. [16], MPC was applied to the operation of a hydrogen-based hybrid energy storage system in a microgrid. A robust optimization model was proposed by Ref. [17] for microgrid operation using ensemble weather forecasts. Another major approach to addressing the challenges of uncertainties is stochastic optimization. For example, a stochastic energy scheduling model in microgrids with intermittent renewable energy resources was proposed in Ref. [18]. A risk-averse stochastic programming method was proposed, which considered not only the expectation but also the variation of the total cost. From the perspective of control objectives, the existing literature focuses on optimizing costs, power system stability such as voltage-frequency control, and environmental objectives such as reducing carbon dioxide emissions. However, the intermittent and nondispatchable nature of RES can pose challenges to system reliability.

Fluctuations in RES output, often due to factors like adverse weather conditions, can lead to periods of unavailability during high electricity demand. ESS are commonly integrated into microgrids or utility grids to mitigate power imbalances. They serve as flexible mediators, capable of storing excess energy and exporting it as needed. Through diverse control strategies, ESS not only provides auxiliary services but also offers financial benefits to end-users. Research on ESS within microgrids explores various applications, including real-time operations for immediate power sharing and frequency regulation, as well as addressing scheduling challenges related to charging strategies and long-term operational optimization [4]. Previous studies have mainly focused on enhancing energy efficiency and operational reliability through the development of energy management systems [19], and the economic implications of real-time ESS operation across different resource, load, and environmental scenarios [20]. However, they often overlook or assume fixed operational costs. Unlike generation resources, short-term dispatch strategies for ESS significantly impact its long-term lifespan. Frequent charging and discharging can degrade battery life [21] while balancing economic and security concerns adds complexity to energy management optimization in microgrids [22]. Expanding ESS capacity enhances operating reserves, thus reducing the risk of load loss but increasing the need for additional capital investment [23]. However, previous studies either neglect or provide rough models for ESS degradation costs. Additionally, to our knowledge, no study considers ESS degradation in an integrated system comprising variable demand and supply.

Given the difficulty of controlling complex systems with multiple inputs and outputs, recent research has heavily focused on employing RL to tackle challenges in microgrid operation [24]. Markov

decision process (MDP) forms the basis of RL theoretical frameworks, enabling the expression of interactive processes through probability theory for problem-solving. For instance, Lu et al. [25] proposed a dynamic electricity pricing algorithm based on artificial intelligence for hierarchical electricity markets, using RL within a layered structure to address pricing trends as a discrete MDP, employing Q-learning [23]. Additionally, Kofinas et al. [26] applied centralized RL in a single-agent system to manage maintenance in solar energy microgrids, optimizing battery consumption and meeting consumer unit service quality requirements. Aaltonen et al. [27] developed an RL solution applicable to multiple time scales, accurately reflecting battery movement through simulation models and bidding on the primary frequency reserves market. Yang et al. [6] introduced an integrated RL algorithm addressing both continuous and discrete problems by combining temporal difference and deterministic policy gradient algorithms, proposing a bidirectional control framework based on Markov for separate demand and supply-side control. While these RL-based microgrid control methods show promise in various environments, they often overlook cooperation and simultaneous control issues among multiple energy sources within the microgrid, as well as convergence difficulties due to environmental complexity. Li et al. [28] introduced a data-driven model predictive control and approximate dynamic programming-based stochastic real-time operating approach for realistic multienergy microgrids, aiming to reduce uncertainty, coordinate multiple energies, and lower operation costs. However, model-based techniques require specialized model creation, posing challenges for practical implementation and necessitating significant historical data and computing power. In contrast, Huang et al. [7] addressed a joint control problem in a multiagent control framework but only considered the supply side, i.e., energy sources as the agents, while neglecting demand-side variations and control. Their approach underestimates the complexity of dealing with continuous and discrete agents, and the critical role of ESS in microgrids is also ignored. Overall, although DRL-based methods can handle complex environments [29,30] and devise optimal control strategies, the results usually depend on the problem formulation and how the system dynamics are handled.

MARL algorithms, such as MADDPG and multiagent proximal policy optimization (MAPPO), have been widely used to control complex environments involving multiple interacting agents. These algorithms are often applied in environments with homogeneous action spaces, where MADDPG is particularly effective for continuous action spaces, and MAPPO is typically used for discrete action spaces. However, real-world applications often involve environments with mixed action spaces, where agents must handle both discrete and continuous actions simultaneously. To address this challenge, several extensions of MARL algorithms have been proposed in the literature. For example, the P-DQN algorithm introduced by Ref. [31] combines deep Q-network (DQN) and deep deterministic policy gradient to handle discrete and continuous action spaces, respectively. However, this approach is designed for single-agent environments, and its applicability is limited in multiagent settings due to the issue of nonstationarity, which arises when agents' policies are not stationary because they are continuously adapting to one another's behaviors. To extend the P-DQN algorithm to multiagent systems, Fu et al. [32] introduced the MAPQN and MAHHQN algorithms, which aim to address nonstationarity by decoupling the decision-making process. Their approach first identifies the optimal discrete action for each agent and then focuses on finding the corresponding continuous action for that specific discrete choice, rather than exploring all possible combinations of discrete-continuous action pairs. This hybrid action space framework allows for more efficient exploration in multiagent environments where agents must handle both types of action spaces.

Similarly, other works, such as Refs. [33–36], have also explored the challenges of MARL with hybrid action spaces. These approaches often involve environments where agents must simultaneously decide on discrete and continuous actions, for example, in gaming scenarios where an agent might need to select a direction and a corresponding speed. These methods handle complex action spaces where each agent must manage both discrete and continuous parameters. In contrast, this work addresses a relatively simpler setting in which only a subset of agents operates with a hybrid action space, and the discrete part is limited to only 0 and 1. In such cases, traditional MARL algorithms like MADDPG are well-suited to handle the environment effectively. As a result, existing solutions designed for complex hybrid action spaces, where each agent must handle both discrete and continuous actions, are not directly applicable here. Given this simpler structure, we have chosen to apply MADDPG for both types of agents (i.e., discrete and continuous), as it offers a more straightforward and computationally efficient approach. This choice allows to leverage MADDPG's strengths in environments with uniform action spaces for each agent, thereby streamlining the training process compared to algorithms tailored for more intricate hybrid settings.

3 System Description

In this article, an integrated microgrid and manufacturing system is considered as illustrated in Fig. 1. The microgrid system includes a solar energy source, wind energy, a generator, and an ESS, also known as batteries. Another energy source is the utility grid. The manufacturing system consists of a serial production line with M machines and $M-1$ buffers. Raw parts arrive at the manufacturing system's source, undergo processing through a series of machines, and the finished products are stored in the sink. A summary of the notations used is provided in Table 1.

The energy requirements within the manufacturing system primarily rely on machine consumption, which is satisfied by one of the available energy sources. In case of excess energy production, the surplus amount is either stored in the ESS or sold back to the utility grid. Our integrated system operates under the following assumptions.

(1) The words "ESS" and "battery" are used interchangeably.
(2) Extra energy generated from the power sources is sold back to the utility grid or used to charge ESS.
(3) To emphasize our focus on the specific production system, the first machine never starves, and the last machine is never blocked.
(4) Any amount of energy can be purchased anytime from the utility grid.
(5) There is at least one source of renewable energy in the microgrid system.

4 Modeling of Dynamic Energy Consumption for the Integrated System

Our integrated system comprises several main components, including a manufacturing system (demand side), an energy generation system (supply side), and ESS. To devise a control mechanism to coordinate all the components for energy-saving purposes, it is essential to derive real-time energy consumption evaluation for each component.

4.1 Manufacturing System. The behavior of the manufacturing system is discussed in detail in our previous study [37]. Nevertheless, we will provide a general overview of the manufacturing system model here. The dynamics of the manufacturing system can be described by the following state-space equation:

$$\dot{\boldsymbol{X}}(t) = \boldsymbol{F}(\boldsymbol{X}(t), \boldsymbol{U}(t), \boldsymbol{W}(t)) \tag{1}$$

where

- $\boldsymbol{X}(t) = [X_1(t), X_2(t), \ldots, X_M(t)]'$ represents the production count of each machine S_i at time t.

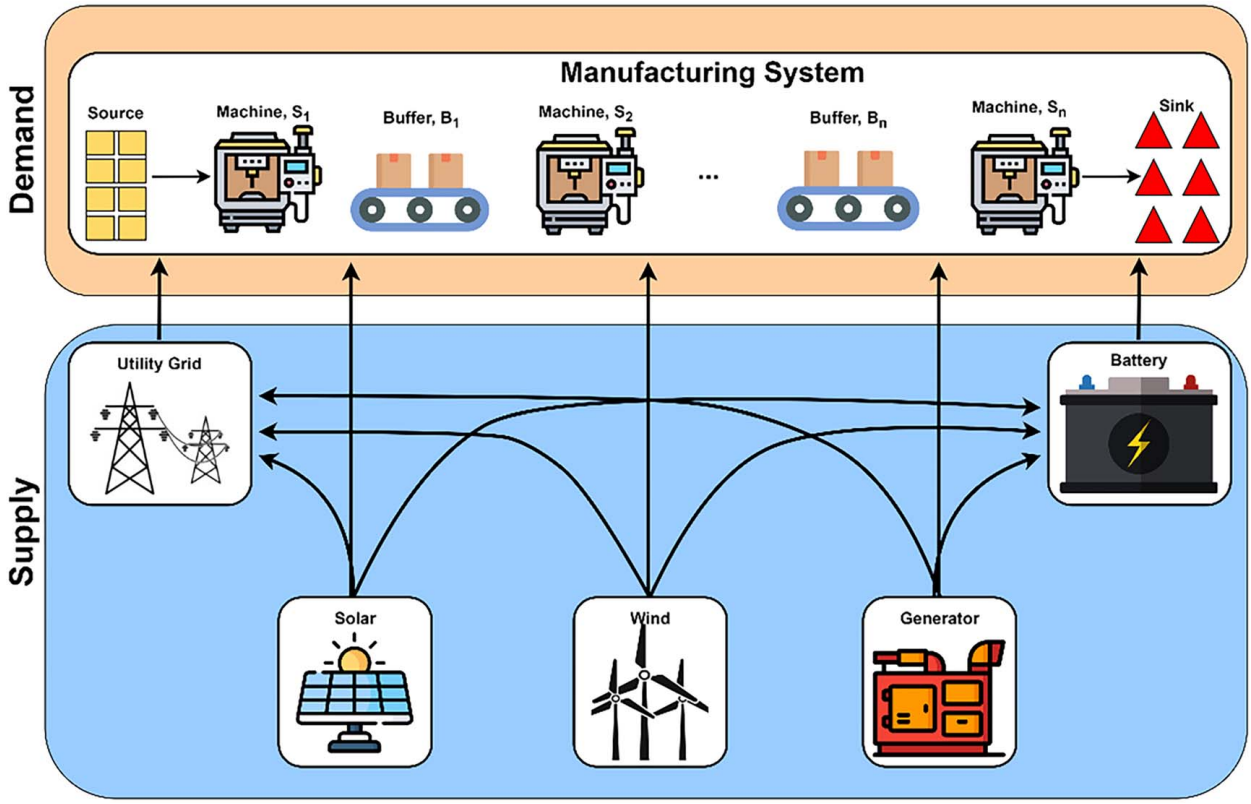

Fig. 1 Integrated manufacturing and microgrid system

- $\boldsymbol{U}(t) = [u_1(t), u_2(t), \ldots, u_M(t)]'$ represents the control inputs at time t. Here, $u_i(t) \in \{0, 1\}$ indicates whether machine S_i is off or on, respectively.
- $\boldsymbol{W}(t) = [w_1(t), w_2(t), \ldots, w_M(t)]'$ represents the disturbances at time t, where $w_j(t)$ describes whether S_j suffers from a disruption at time t. If there exists $\vec{e}_k \in \boldsymbol{E}$ such that $\vec{e}_k = (i, t_k, d_k)$, $t \in [t_k, t_k + d_k]$ and $\boldsymbol{E}$ is a set of disruption events, then, $w_i(t) = 1$, otherwise, $w_i(t) = 0$. Define the status of a machine S_i at time t as $\theta_i(t)$, i.e., $\theta_i(t) = 1 - w_i(t)$. A machine S_i is up at time t when $w_i(t) = 0$, and down when $w_i(t) = 1$.

The accumulated production difference between two machines S_i and S_j, $i, j \in 1, 2, \ldots, M, i \neq j$, within the time period $[0, t]$, follows the conservation of flow, which could be represented with the following equations:

$$X_i(t) - X_j(t) = \tau_{ij}(t) = \begin{cases} \sum_{k=j+1}^{i} b_k(0) - \sum_{k=j+1}^{i} b_k(t), & i > j \\ \sum_{k=i+1}^{j} b_k(t) - \sum_{k=i+1}^{j} b_k(0), & i < j \end{cases} \tag{2}$$

The production difference $\tau_{ij}(t)$ is bounded by the condition that all buffers between machines S_i and S_j are full (for $i < j$) or empty (for $i > j$). Denote the boundary as $\beta_{ij}(t)$.

$$\beta_{ij}(t) = \begin{cases} \sum_{k=j+1}^{i} b_k(0), & i > j \\ \sum_{k=i+1}^{j} B_k - \sum_{k=i+1}^{j} b_k(0), & i < j \end{cases} \tag{3}$$

Thus, $\tau_{ij}(t) \leq \beta_{ij}(t)$. Considering the interactions between S_i and S_j, in the case of $\tau_{ij}(t) < \beta_{ij}(t)$, machine S_i is not starved or blocked by S_j; thus, it will process parts at its own rated speed. If $\tau_{ij}(t) = \beta_{ij}(t)$, the processing speed of machine S_i will be constrained by machine S_j. Define a segment function $\xi(u, v)$ as

$$\xi(u, v) = \begin{cases} +\infty, & u < 0 \\ v, & u = 0 \end{cases}$$

Then, the actual process speed of machine S_i can be described as

$$\dot{X}_i(t) = \min\left\{\frac{\xi((X_i(t) - X_j(t)) - \beta_{ij}, u_j(t)(1 - W_j(t)))}{T_j(t)}, \frac{u_i(t)(1 - W_i(t))}{T_i(t)}\right\} \tag{4}$$

Extending this equation to all machines in the system, we have

$$\dot{X}_i(t) = \min\left\{\begin{array}{c} \frac{\xi((X_i(t) - X_1(t)) - \beta_{i1}), u_1(t)(1 - W_1(t))}{T_1(t)} \\ \frac{\xi((X_i(t) - X_2(t)) - \beta_{i2}), u_2(t)(1 - W_2(t))}{T_2(t)} \\ \vdots \\ \frac{u_i(t)(1 - W_i(t))}{T_i(t)} \\ \vdots \\ \frac{\xi((X_i(t) - X_M(t)) - \beta_{iM}), u_M(t)(1 - W_M(t))}{T_M(t)} \end{array}\right\} \tag{5}$$
$$= f_i(\boldsymbol{X}(t), \boldsymbol{U}(t), \boldsymbol{W}(t))$$

Thus, the state-space function of the production count could be summarized as

$$\boldsymbol{X}(t) = \begin{bmatrix} F_1(\boldsymbol{X}(t), \boldsymbol{U}(t), \boldsymbol{W}(t)) \\ \cdot \\ \cdot \\ \cdot \\ F_M(\boldsymbol{X}(t), \boldsymbol{U}(t), \boldsymbol{W}(t)) \end{bmatrix} = \boldsymbol{F}(\boldsymbol{X}(t), \boldsymbol{U}(t), \boldsymbol{W}(t)) \tag{6}$$

Table 1 Table of notations

Symbol	Definition
A_d	Discrete actions
a_i	Action of machine i
A_c	Continuous action
EU	Energy cost for purchased energy from utility grid
MC	Operation cost of the microgrid
TP	Throughput reward
SB	Reward for selling back energy
P	Energy usage from utility grid
r^b	Electricity cost per unit purchased from utility grid
P^m	Energy purchased from grid for manufacturing
P^b	Energy purchased from grid for battery charging
s^m	Solar energy for manufacturing
w^m	Wind energy for manufacturing
g^m	Generator energy for manufacturing
b^m	Energy discharged from the battery supporting manufacturing
E^{mfg}	Total energy consumed by manufacturing system at time-step t
PC_{it}	The amount of power drawn from machine i at time t
Δt	Time between two consecutive time-steps
r^s_{omc}	Unit operation and maintenance cost of solar energy
r^w_{omc}	Unit operation and maintenance cost of wind energy
r^g_{omc}	Unit operation and maintenance cost of generator energy
e^s	Energy generated from the solar
e^w	Energy generated from the wind
e^g	Energy generated from the generator
$C_B(t)$	Battery degradation cost at time t
C	Number of charge/discharge cycles
D	Depth of discharge
P_B	Battery operating mode signal (charging/discharging signal)
ΔT	Summation of k intervals
E_A	Actual capacity of the battery (MWh)
$C_{\text{BC}}(t, D(\Delta T))$	Battery degradation cost for one charging–discharging cycle
C_{cap}	Capital cost of the battery per unit energy
$E_{B,r}$	Rated energy capacity of the battery
$f(t)$	Battery's charging/discharging cycle transition identification function
$\text{AE}(t)$	Accumulated energy (kWh) related to one charging or discharging cycle
PC	Production count at time t
E^{sb}	Total sold back energy
r^{sb}	Reward for one unit of sold back energy
SOC	State of the charge of battery
r^p	Reward associated with each finished product

$$\boldsymbol{Y}(t) = X_M(t) = [0 \ldots 0\ 1]\boldsymbol{X}(t) = \boldsymbol{H}(\boldsymbol{X}(t)) \tag{7}$$

where $\boldsymbol{Y}(t)$ is the production output of the last machine S_M at time t.

The buffer levels of a certain buffer b_{i+1} at time t could be calculated as

$$b_{i+1}(t) = X_i(t) - X_{i+1}(t) + b_{i+1}(0) \tag{8}$$

If the sensor data of random disruption events $\boldsymbol{E}$, the machine inputs $u_1(0)$, $u_2(0)$,…, $u_M(0)$, and the initial buffer levels $b_1(0)$, $b_2(0)$,…, $b_{M-1}(0)$ are provided, the detailed system state at any given time can be recursively evaluated by (2)–(8).

Each product is associated with a unit reward r^p. We denote the total reward from production throughput as TP, which can be calculated as follows:

$$\text{TP} = Y(t) \times r^p \tag{9}$$

At each time t, an energy demand originates from the manufacturing system which is the collective power usage of all the operating machines. The energy consumed by the manufacturing system E^{mfg} can be defined as

$$E^{\text{mfg}}(t) = \sum_{i=1}^{N} \text{PC}_{it} \times \Delta t \tag{10}$$

where PC_{it} is the amount of power drawn from machine i during t to $t+1$.

4.2 Energy System. The energy system is composed of several elements: solar power source, wind power source, generator power source, and utility grid, which are discussed in the following subsections.

4.2.1 Microgrid System. The microgrid encompasses various components that collectively determine its total operational cost [6]. In our scenario, these components include solar panels, wind turbines, a generator, and a battery system. The total operational cost of the microgrid at each time-step t denoted as $\text{MC}(t)$ can be computed as follows:

$$\text{MC}(t) = e^s(t) \times r^s_{\text{omc}} + e^w(t) \times r^w_{\text{omc}} + e^g(t) \times r^g_{\text{omc}} + C_B(\text{t}) \tag{11}$$

where $e^s(t)$, $e^w(t)$, and $e^g(t)$ represent, respectively, the energy generated by solar, wind, and generators; r^s_{omc}, r^w_{omc}, and r^g_{omc} represent, respectively, the rated operating and maintenance cost of solar, wind, and generator power source; $C_B(t)$ is the cost of battery degradation at time-step t. Each component of the microgrid is explained below.

Energy sold back: The extra energy generated by the different renewable power sources is sold back to the utility grid. Let us denote the energy sold back as SB, which can be calculated as

$$\text{SB}(t) = E^{\text{sb}}(t) \times r^{\text{sb}} \tag{12}$$

where $E^{\text{sb}}(t) = s^{\text{sb}}(t) + w^{\text{sb}}(t) + g^{\text{sb}}(t)$, and r^{sb} is the unit reward from sold back energy.

Solar power: The energy generated from solar power source $e^s(t)$ at any time-step t typically depends on the solar irradiation $I(t)$ at the corresponding time t and the size of the solar panels. There is also an operation and maintenance cost associated with solar panels denoted as r^s_{omc}. Based on Ref. [38], the solar energy $e^s(t)$ can be calculated as

$$e^s(t) = \begin{cases} 0, & \text{if } a^s(t) = 0 \\ I(t) \times A \times \delta \times \Delta t/1000, & \text{if } a^s(t) = 1 \end{cases} \tag{13}$$

where $a^s(t)$ denote the switch status of solar at time-step t, which can be on or off. $I(t)$ is the solar irradiance in W/m^2, A is the area of the solar panel, and δ is the efficiency of the solar PV system.

Wind power: The energy generated from wind depends on the wind parameters and can be calculated as [6]

$$e^w(t) = \begin{cases} 0, & \text{if } a^w(t) = 0, \text{ or } W(t) \langle W_{\text{ci}} \text{ or } W(t) \rangle W_{\text{co}} \\ n_w R_w \Delta t, & \text{if } a^w(t) = 1 \text{ and } W^r \le W(t) < W_{\text{co}} \\ n_w R_w \dfrac{W(t) - W_{\text{ci}}}{W^r - W_{\text{ci}}} \Delta t, & \text{if } a^w(t) = 0 \text{ and } W_{\text{ci}} \le W(t) < W^r \end{cases} \tag{14}$$

where $a^w(t)$ represents the switch status of the wind power source, n_w is the number of wind turbines, $W(t)$ is the wind speed at time-step t, W_{ci} is the cut-in speed while W_{co} is the cutoff speed, W^r represents the rated wind speed, and R_w denotes the rated power of wind turbine (kW). The rated power can be calculated as

$$R_w = \frac{1}{2} \times \rho \times \pi \times r_w^2 \times v_{\text{avg}}^3 \times \theta \times \eta_t \times \eta_g / 1000 \tag{15}$$

where ρ denotes the density of air, r_w is the radius of the wind turbine blade, v_{avg} is average wind speed, θ is the power coefficient, and η_t and η_g represent the gearbox transmission efficiency and electrical generator efficiency, respectively.

The wind speed $W(t)$ and the solar irradiance $I(t)$ affect the system dynamics and cost function but are unaffected by the control actions. They depend on time and weather conditions. The model formulation assumes access to a deterministic forecast of this information. Both $W(t)$ and $I(t)$ are derived from data spanning 1 year, assumed to consist of 360 days with 24 h each, totaling 8640 h [39,40].

Generator power: The energy from the generator source $e^g(t)$ is a nonrenewable source of energy that can be utilized in case of excessive demands or unavailability of renewable energies [6]. It can be calculated as

$$e^g(t) = \begin{cases} 0, & \text{if } a^g(t) = 0 \\ n_g \times R_g \times \Delta t, & \text{if } a^g(t) = 1 \end{cases} \tag{16}$$

where n_g is the number of generators, $a^g(t)$ is the status of the generator at time t, and R_g is the rated power of the generator (kW).

Energy storage system: The ESS is modeled based on Ref. [41]. The battery's lifespan, measured in the number of charge–discharge cycles C, varying inversely with the depth of discharge D [42], can be represented as

$$C(D) = \beta_0 \times D^{-\beta_1} \times \exp(-\beta_2 \times D) \tag{17}$$

where β_0, β_1, and β_2 are positive curve fitting coefficients. This expression is valid for different types of batteries with different parameters [43].

Continuous monitoring of the battery operating mode signal, denoted as P_B, enables the identification of charging or discharging events. A charging cycle is indicated when $P_B(t)$ is positive, indicating power consumption, while a discharging cycle is indicated when $P_B(t)$ is negative, representing power delivery. Each charging or discharging cycle can last for k intervals.

Based on Ref. [42], the depth of discharge D after a discharging or charging event during time ΔT can be represented as

$$D(\Delta T) = \left| \frac{P_B(t)\eta\Delta T}{E_A(t)} \right| \tag{18}$$

where $P_B(t)$ is the charging or discharging power, $E_A(t)$ is the actual capacity at time t, and η is the charging or discharging efficiency. It can be represented as

$$\eta = \begin{cases} \eta_c, & \text{if } P_B(t) \geq 0 \\ \dfrac{1}{\eta_d}, & \text{if } P_B(t) < 0 \end{cases} \tag{19}$$

The amount of life lost due to the preceding charging or discharging cycles is equivalent to $1/2\ C(D(t))$ [43,44]. Therefore, the battery degradation cost, denoted as $C_{\text{BC}}(t, D(\Delta T))$, for one charging or discharging cycle can be written as

$$C_{\text{BC}}(t, D(\Delta T)) = \frac{C_{\text{cap}} \times E_{B,r}}{2C(D(\Delta T))} \tag{20}$$

where C_{cap} indicates the capital cost of the battery per unit energy and $E_{B,r}$ is the rated energy capacity of the battery. The actual capacity of the battery decreases after the completion of a charging or discharging cycle. This degradation analogizes $1/2\ C(D)$. Therefore, the actual capacity of the battery after completion of a charging or discharging cycle can be calculated as

$$E_A(t + \Delta t) = E_A(t) - \frac{E_{B,r}}{2C(D(\Delta T))} \tag{21}$$

Let $f(t)$ be a binary variable to identify the state transition during charging and discharging events in two consecutive time intervals, which can be expressed as

$$f(t) = \begin{cases} 1, & \text{if } P_B(t)P_B(t-1) \leq 0 \\ 0, & \text{if } P_B(t)P_B(t-1) > 0 \end{cases} \tag{22}$$

We define another variable, AE, as the accumulated energy (kWh) related to one charging or discharging cycle. It can be written as

$$\text{AE}(t) = (1 - f(t))\text{AE}(t-1) + P_B(t)\Delta t \tag{23}$$

Hence, the cost of battery degradation $C_B(t)$ over successive time intervals can be represented by the state transition signal $f(t)$ in conjunction with the accumulated energy AE(t) as follows:

$$C_B(t) = C_{\text{BC}}\left(t, \frac{\text{AE}(t)}{E_A(t)}\right) - (1 - f(t))C_{\text{BC}}\left(t, \frac{\text{AE}(t-1)}{E_A(t-1)}\right) \tag{24}$$

4.2.2 Utility Grid. Typically, a utility grid is utilized as a last resort when no other power source is available or when it offers a more cost-effective option. The cost of power usage from the public grid, denoted as EU, can be defined as follows:

$$\text{EU} = P(t) \times r^b(t) \tag{25}$$

where $P(t)$ is the energy usage from utility grid while $r^b(t)$ is the electricity cost per unit. The purchased power is mainly used to support the manufacturing system and charge the battery, so $P(t)$ can be written as

$$P(t) = (P^m(t) + P^b(t)) \tag{26}$$

where $P^m(t)$ is the energy purchased from the grid to support the manufacturing and $P^b(t)$ is the energy purchased to charge the battery.

$$P^m(t) = E^{\text{mfg}}(t) - (s^m(t) + w^m(t) + g^m(t) + b^m(t)) \tag{27}$$

where $E^{\text{mfg}}(t)$ is the total energy consumed by the manufacturing system at time t as given in Eq. (10).

Our proposed integrated system is based on the following constraints.

(1) ESS in the microgrid should maintain a certain state of charge (SOC), i.e.,

$$\begin{aligned} \text{SOC}_{\min} \leq\ & \text{SOC}(t) + (s^b(t) + w^b(t) \\ & + g^b(t) + p^b(t))\eta - \frac{b^m(t)}{\eta} \leq \text{SOC}_{\max} \end{aligned} \tag{28}$$

where $\text{SOC}_{\min}$ and $\text{SOC}_{\max}$ are, respectively, the minimum and maximum thresholds, which cannot be exceeded by the battery charge. They depend on the actual capacity of the battery $E_A(t)$ and are given as

$$\text{SOC}_{\min} = 0.05 \times E_A(t) \tag{29}$$

$$\text{SOC}_{\max} = 0.95 \times E_A(t) \tag{30}$$

(2) The amount of energy provided by solar to the manufacturing, battery storage, and sold back cannot exceed the total amount of energy generated by the solar panels, i.e.,

$$s^m(t) + s^b(t) + s^{\text{sb}}(t) = e^s(t) \tag{31}$$

Similarly, for the wind and generator energies, the energy generated cannot exceed the energy supplied which can be presented as follows:

$$w^m(t) + w^b(t) + w^{\text{sb}}(t) = e^w(t) \tag{32}$$

$$g^m(t) + g^b(t) + g^{\text{sb}}(t) = e^g(t) \tag{33}$$

(3) The battery cannot be charged and discharged simultaneously, i.e.,

$$(s^b(t) + w^b(t) + g^b(t) + p^b(t)) \times b^m(t) = 0 \tag{34}$$

(4) The energy cannot be purchased from the grid and sold back to the grid simultaneously, i.e.,

$$(s^{\text{sb}}(t) + w^{\text{sb}}(t) + g^{\text{sb}}(t))(p^m(t) + p^b(t)) = 0 \tag{35}$$

(5) The energy purchased from the grid can be used for either supporting manufacturing or charging the battery, but not simultaneously, i.e.,

$$p^m(t) \times p^b(t) = 0 \tag{36}$$

(6) The total energy consumed by manufacturing must be equal to the total energy supplied by different sources to the manufacturing as given by Eq. (27).

5 Control Problem Formulation and Solution

In this integrated system, the control objective is twofold: to enhance the throughput while conserving energy. Control actions on the demand side entail activating and deactivating machines. Conversely, on the supply side, actions involve activating/deactivating energy sources and allocating energy from each source to the manufacturing system, battery storage, and selling excess energy back to the utility grid. Managing such an integrated control setup is exceptionally challenging due to the stochastic and nonlinear nature of the manufacturing system, which lacks closed-form representation [37,45]. Furthermore, controlling the supply side exacerbates the complexity, constituting an NP-hard problem. Real-time execution further compounds the challenge. The absence of a closed-form representation impedes the application of traditional control methods. Consequently, we approach the problem as a Dec-POMDP and tackle it using MARL. The overarching aim is to determine an optimal policy in stochastic scenarios, mapping states to actions, to maximize rewards.

5.1 Dec-POMDP Framework for MARL. In the MARL framework, at state s_t, each agent $i, i = 1, 2, \ldots, n$, selects their own action a_t^i and receives the corresponding reward $r_t^i(s_t, a_t^1, a_t^2, \ldots, a_t^n)$. These actions result in a joint action $\boldsymbol{a_t} = (a_t^1, a_t^2, \ldots, a_t^n)$, and the state transitions to the next state s_{t+1} with the transition probability $p(s_{t+1}|s_t, \boldsymbol{a_t})$ such that it satisfies $\sum_{s_{t+1} \in S} p(s_{t+1}|s_t, \boldsymbol{a_t}) = 1$. Each agent strives to maximize their own discounted accumulated reward by selecting actions a_t^i under the policy π_t^i, which represents the decision-making rule for the distribution of their actions. The policy for each agent i is represented by $\pi^i = (\pi_0^i, \pi_1^i, \ldots, \pi_t^i, \ldots)$. Given an initial state s, and discount factor $\gamma \in [0, 1)$, the accumulated reward can be expressed as

$$v^i(s, \pi^1, \pi^2, \ldots, \pi^n) = \sum_{t=0}^{\infty} \gamma^t E(r_t^i | \pi^1, \pi^2, \ldots, \pi^n, s_0 = s) \tag{37}$$

Since the transition probabilities are unknown in this system, a model-free MARL algorithm is employed. This approach eliminates the need for assumed transition probabilities and focuses solely on defining observations, actions, and rewards.

5.1.1 Observations. Observations refer to information that each agent has about the current state of the system. These "raw observations" only reflect local information from individual agents and may result in sub-optimal performance in coordinated control problems. Since our system comprises both continuous agents, i.e., energy sources, and discrete agents, i.e., machines, each type has a different observation. The ith discrete agents' observation is represented as

$$o^{d_i}(t) = \{\theta^i(t), b^1(t), b^2(t), \ldots, b^{M-1}(t)\} \tag{38}$$

where $\theta^i(t)$ represents the status of machine i at time t, which can be either off or on (represented as 0 and 1, respectively), and b_t^i represent the level of buffer i at time t.

Similarly, the ith continuous agent's observation can be represented as

$$o^{c_i}(t) = \{\text{SOC}(t), \mathbf{C}_i\} \tag{39}$$

where SOC(t) represents the state of charge of the ESS at time t and $\mathbf{C}_i = [c_m, c_b, c_{\text{sb}}]$ with c_m being the amount of energy dispatched to manufacturing, c_b being the amount of energy dispatched to the battery, and c_{sb} being the amount of energy sold back to the utility grid. In this case, $i \in \{\text{solar,wind,generator}\}$ and the values from solar and wind depend on the corresponding values of the solar irradiation and wind speed.

In a Dec-POMDP, the state usually comprises the observations of all agents, i.e.,

$$s_t = \{o^{d_i}(t), o^{c_i}(t)\} \tag{40}$$

5.1.2 Action. The actions are also distributed based on the type of agent. A discrete agent, i.e., the machine can either be on or off. It is defined as

$$A_d = \{a_i\} \tag{41}$$

where $a_i \in \{0, 1\}$ represents the action of machine i to be turned on or off.

Similarly, the continuous agents including solar, wind, and generator energy sources are responsible for deciding the proportion of source to manufacturing, battery storage, and sold back. It is defined as

$$A_c = \{a_\theta^j, a_m^j, a_b^j, a_{\text{sb}}^j\} \tag{42}$$

where $a_\theta^j \in [0, 1]$ represents the status of the power source j, which can be either switched on or off; a_m^j represents the proportion of power from source j to the manufacturing; a_b^j represents the proportion to battery; and a_{sb}^j is the proportion sold back to the utility grid. In this case, $j \in \{\text{solar,wind,generator}\}$.

5.1.3 Reward. The global reward encapsulates the effectiveness of the collective actions undertaken by all agents. In the context of this article, our aim is to simultaneously minimize the total energy cost and maximize production throughput. As previously mentioned, energy costs stem from various sources: energy purchased from the utility grid, operational expenses of the microgrid responsible for energy generation, and revenue generated by selling excess energy back to the utility grid, which we seek to optimize. Likewise, production throughput originates from the manufacturing system and is contingent upon machine operations. With these objectives in mind, the following reward function is formulated:

$$r_i = -\text{EU} - \text{MC} + \text{TP} + \text{SB}_i \tag{43}$$

where EU is the power purchased from the utility grid, given in Eq. (25); MC is the operational cost of the microgrid given in Eq. (11); TP is the manufacturing system's throughput as given in Eq. (9); and SB is the reward from energy sold back to the utility grid by agent i at time t, given in Eq. (12).

5.2 MADDPG. We utilize the MADDPG algorithm [46] to address our control problem. Each machine, denoted as S_i, in the manufacturing system, and each power source, solar, wind, or generator, is regarded as an individual agent. Given that machines are simply switched on or off, discrete actions suffice for their control. However, energy sources must determine the distribution of energy among manufacturing, battery storage, and selling back, necessitating continuous action values. Moreover, energy

sources must decide whether to activate or deactivate each source. Therefore, the MADDPG algorithm is employed to handle both discrete and continuous types of agents.

Figure 2 provides a general overview of the control structure operating in the integrated microgrid-manufacturing system. Each agent is initialized with its actor and critic networks. Agents delineated by solid boundary lines represent continuous agents, while those delineated by dashed lines represent discrete agents. Actions are initialized based on initial observations. Actions from continuous agents depend on additional parameters; for instance, solar energy is contingent upon solar irradiance data. The available solar power is then distributed based on actions generated by the solar power source. Similarly, wind energy depends on wind speed, and generator energy depends on the number of generators and fuel price. For discrete actions, the decision is solely to turn machines on or off. Based on these actions, total energy demand is generated depending on the operational status of all machines. This demand is then compared with the total available energy from different power sources for manufacturing. If the demand is less than available energy, it is met, and any surplus energy may be used for battery charging or sold back to the utility grid. Conversely, if the demand exceeds energy supply, battery storage may be utilized, or additional energy purchased from the utility grid. Using Eq. (43), the reward is calculated and sent to agents along with corresponding observations. The control algorithm operates in this manner, seeking an optimal policy to maximize total reward while minimizing energy consumption.

Algorithm 1. Pseudocode for the proposed MADDPG with continuous and discrete agents

Inputs:
- Manufacturing environment with continuous and discrete action spaces
- Number of agents
- Continuous policy network (μ) for continuous actions
- Discrete policy network (π) for discrete actions (with Softmax layer)
- Continuous target network (μ')
- Discrete target network (π')
- Critic network (Q) for evaluating actions
- Critic target network (Q')
- Replay buffer
- Hyperparameters (learning rates, discount factor, exploration noise, etc.)

Initialize:
- For each agent:
  - Continuous policy network (μ): Neural network for continuous actions
  - Discrete policy network (π): Neural network for discrete actions with Softmax
  - Continuous target network (μ'): Copy of the continuous policy network
  - Discrete target network (π'): Copy of the discrete policy network
  - Critic network (Q): Neural network for evaluating action-value function
  - Critic target network (Q'): Copy of the critic network
  - Replay buffer: Store experiences (S_t, A_c, A_d, R_t, S_{t+1})
  - Optimizers for continuous and discrete networks

For each episode:
Initialize the environment and get the initial states (s_1, s_2, …, s_N)
For each time-step within the episode:
For each agent:
- Compute continuous action A_c using μ
- Compute action probabilities for discrete actions using π with Softmax layer
- Sample discrete action A_d based on these probabilities
- Combine A_c and A_d to form the complete action vector
- Execute the combined action in the environment
- Observe new states and reward
- Store experience (S, A_c, A_d, R_t, S') in replay buffer

Sample a batch of experiences from replay buffer
For each agent:
- Compute the target continuous action using target network (μ'): $A'_c = \mu'(s')$
- Compute the target discrete action using target network (π'): $A'_d = \pi'(s')$
- Compute the target Q value: $Q_{\text{target}} = r + \gamma \times Q'(s', A'_c, A'_d)$
- Update critic network (Q):
  - Compute critic loss: $\text{loss}_Q = L(Q(s, A_c, A_d), Q_{\text{target}})$
  - Optimize critic network with loss_Q
- Update continuous policy network (μ) by minimizing the loss between Q value and Q_{target}:$\text{loss}_c = L(Q(s, A_c, A_d), Q_{\text{target}})$
Optimize continuous policy network with loss_c
- Update discrete policy network (π) by maximizing the expected reward based on action probabilities: $\text{loss}_d = -L(Q(s, A_c, A_d), Q_{\text{target}})$
Optimize discrete policy network with loss_d using A2C
Update target networks (μ' and π') using Polyak averaging
$\mu' \leftarrow \tau\mu + (1-\tau)\mu'$
$\pi' \leftarrow \tau\pi + (1-\tau)\pi'$
$Q' \leftarrow \tau Q + (1-\tau)Q'$

The MADDPG framework is extended to handle both continuous and discrete agents. To achieve this, two distinct policy networks have been integrated: a continuous policy network, denoted as μ, and a discrete policy network, denoted as π. The continuous policy network μ is designed to generate continuous actions A_c, while the discrete policy network π produces probabilities for discrete actions A_d. The discrete actions are sampled based on these probabilities, which are computed using a Softmax layer in π.

The critic network, parameterized by ψ, evaluates state-action pairs (S,A_c,A_d) and is used to compute the Q values. To improve stability and performance, target networks are employed for both continuous and discrete agents. The target continuous network μ' and the target discrete network π' are copies of the continuous and discrete policy networks, respectively. These target networks are used to compute the target Q values, which guide the updates of the critic network and policy networks.

During each episode, the environment is initialized, and initial states are obtained. For each time-step t within an episode, each agent computes its continuous action A_c using the continuous policy network μ. Simultaneously, the discrete policy network π computes action probabilities for discrete actions, from which discrete actions A_d are sampled. The combined actions A_c and A_d are executed in the environment, and new states and rewards are observed. This experience, consisting of the state, actions, reward, and next state, is stored in the replay buffer. A batch of experiences is then sampled from the replay buffer for training. For each agent, the target continuous action A'_c and the target discrete action A'_d are computed using the target networks μ' and π'. The target Q value is calculated as $Q_{\text{target}} = R_t + \gamma Q(S', A'_c, A'_d; \psi')$, where γ is the discount factor. The critic network is updated by minimizing the loss function, which measures the difference between the predicted Q values and the target Q values. The loss function is given by $\text{loss}_Q = \mathbb{E}_{(S, A_c, A_d, R_t, S') \sim \text{Replay Buffer}}[(Q(S,A_c,A_d, ; \psi) - Q_{\text{target}})^2]$. For the continuous policy network μ, the policy loss is computed using the critic network. The loss function for the continuous actions is given by $\text{loss}_c = -\mathbb{E}_{(S,A_c,A_d,R_t,S') \sim \text{Replay Buffer}}[L(Q(S, A_c, A_d, ; \psi), Q_{\text{target}})]$, and the network is optimized accordingly.

Similarly, the discrete policy network π is updated using its own policy loss function, which is given by $\text{loss}_d = -\mathbb{E}_{(S,A_c,A_d, R_t,S') \sim \text{Replay Buffer}}[L(Q(S,A_c,A_d,; \psi),Q_{\text{target}})]$. This loss is optimized using advantage actor critic (A2C). Finally, the target networks are updated using Polyak averaging to ensure stability in the training process. The update rules for the target networks are: $\mu' \leftarrow \tau\mu + (1-\tau)\mu'$ and $\pi' \leftarrow \tau\pi + (1-\tau)\pi'$, where τ is the update rate. This approach effectively integrates both continuous and discrete action handling within the modified MADDPG

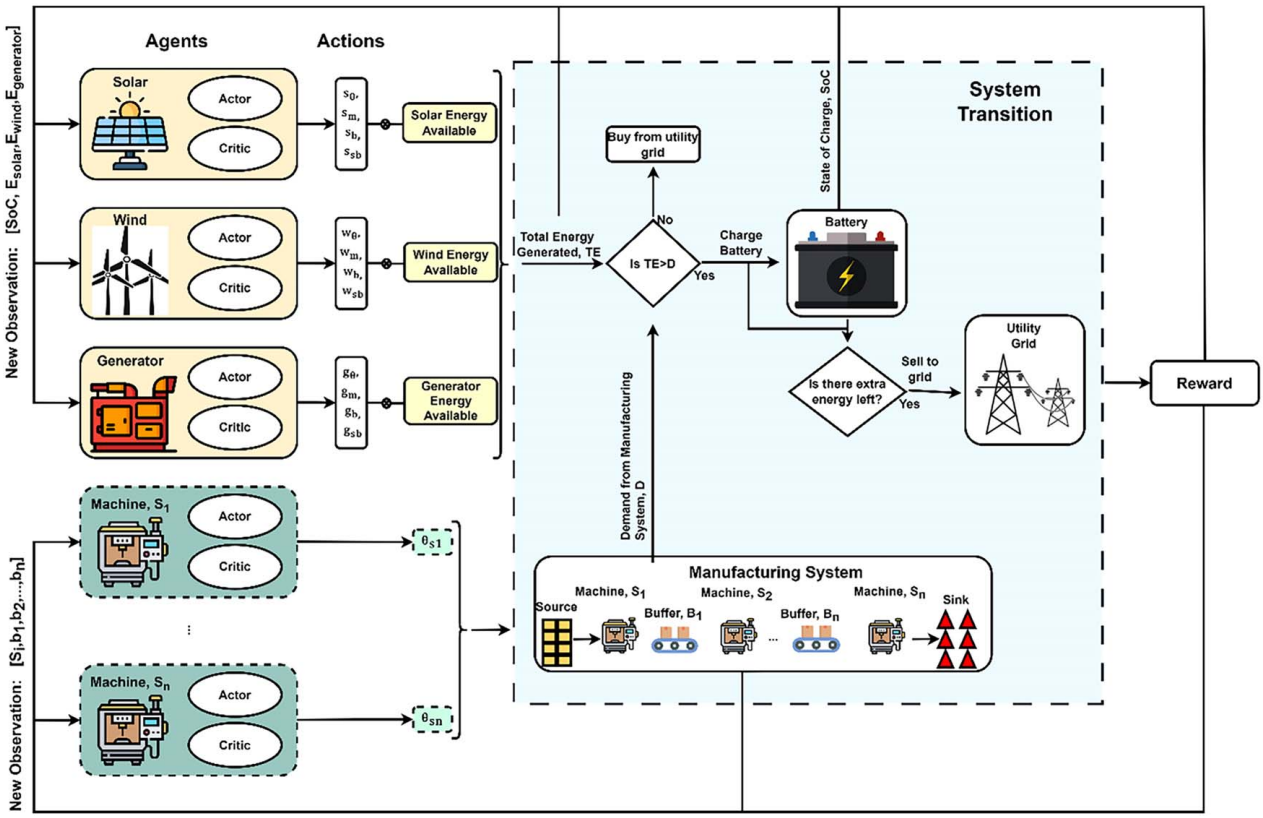

Fig. 2 MADDPG-based control framework for the integrated system

framework, providing a robust method for managing complex environments with mixed agents.

6 Case Study

This section conducts numerical studies to validate the effectiveness of the proposed modeling and multiagent control solutions. An integrated microgrid-manufacturing system is examined to demonstrate the method. The manufacturing data is based on automobile production lines located in the Midwest of the United States and it is assumed that the production operates on three shifts. The manufacturing system comprises five machines and four buffers, with system parameters obtained from Refs. [6,30]. For all the machines, random disruptions are generated, assuming a geometric distribution using the mean time between failures and the mean time to repair. For comparison, random and centralized RL-based control approaches are employed as alternative methods to address the same problem. Various performance metrics are considered, including (1) system production throughput using the trained policy, (2) microgrid total operational cost, (3) amount of energy sold back to the utility grid, (4) ESS degradation cost, and (5) total energy cost. The case study aims to arrive at several significant conclusions: (1) the proposed MARL control scheme effectively optimizes energy distribution and management and (2) the proposed MARL scheme and a rule-based policy derived from the MARL policy surpasses a centralized RL-based policy and random control methods in terms of the performance metrics such as the throughput and the total energy cost.

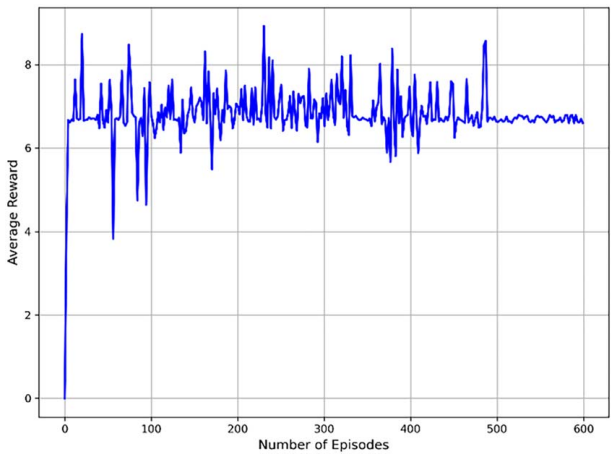

Fig. 3 Training curve: reward versus time

6.1 Experiment Parameter Setting. The agents in this study are trained using the MADDPG algorithm over approximately 600 episodes. Each episode encompasses 3000 time-steps, which simulates a duration of 1 week, considering a work schedule of 5 days and 10 h per day. During training, after the completion of each episode, the environment is reset, and agents are randomly initialized to ensure varied conditions. Training is conducted until convergence is reached, which typically occurs around 500 episodes, as illustrated by the reward training curve shown in Fig. 3.

The neural network architecture used for training consists of two fully connected hidden layers, each containing 64 hidden units. The reward function is derived based on the formulation provided in Eq. (43), with a discount factor (γ) set at 0.99. The actor network's learning rate (α_a) is set to 0.003, while the critic network's learning rate (α_c) is 0.001.

The parameters for the microgrid used in the experiment are sized according to the manufacturing load methods outlined in Ref. [11]. Details for the wind turbine, battery storage system, solar panel, and generator are provided in Table 2. Solar irradiance and wind speed data are sourced from Solar Energy Local [39] and the State Climatologist of Illinois [40], respectively.

To avoid computational overflow, all parameters are scaled by employing alternative units of measurement. Specifically, distance is measured in kilometers (1 km = 1000 m), time in hours (1 h = 3600 s), speed in kilometers per hour (1 km/h = 3.6 m/s), energy in megawatts (1 MW = 1,000,000 W), monetary cost in units of $10,000, area in square kilometers (1 km^2 = 1,000,000 m^2), and mass in millions of kilograms (1,000,000 kg). The time period is recorded in hours.

6.2 Effective Evaluation of the Proposed Method. Following the training of the algorithm, the proposed integrated control policy is assessed using performance metrics. The integrated system operates for a duration of 100 time-steps (hours), after which various performance metrics are observed and analyzed.

6.2.1 Total Energy Cost. Reducing overall energy consumption stands as a key objective of this study. As mentioned earlier, it encompasses the costs associated with microgrid operations, energy procurement from the utility grid, and energy sold back. To assess the efficacy of our MARL-based control scheme, we compare the total energy consumption under the MARL policy with that of centralized RL-based control, random control, and rule-based control derived from the trained MARL policy. The rule-based policy is explained in the subsequent section.

As depicted in Fig. 4, random control exhibits the highest energy consumption, given its lack of adherence to specific rules. Among the remaining three control methods, RL-based control consumes more energy than both MARL-based and rule-based controls, The rule-based approach closely resembles the MARL policy it is derived from. Overall, MARL-based control incurs the lowest energy costs due to its training to manage interactions and uncertainties among different agents. Over a 100-h period, MARL-based control results in an energy expenditure of approximately $5500. Rule-based control, which is summarized from trained MARL-based policy and will be introduced in the next section, costs around $7500, RL-based control around $9000, and random control approximately $19,000.

6.2.2 Throughput. Another performance metric involves maximizing the system throughput, defined as the number of parts produced by the last machine S_M of the manufacturing system. The throughputs are compared between the RL-based control policies and the random control strategies. Figure 5 illustrates the comparison of throughputs, demonstrating a notable increase in the throughput with multiagent control compared to the other methods.

Table 2 Parameters for the microgrid

Parameters	Value
W_{ci} (m/s)—cut-in speed	3
W_{co} (m/s)—cutoff speed	11
W^r(m/s) —rate speed	7
ρ (kg/m^3)—air density	1.225
r (m)—blade radius	25
η_t—gear box efficiency	0.9
η_g—generator efficiency	0.9
θ—power coefficient	0.593
r^w_{omc}($/kWh)—unit cost wind	0.08
n_w—number of wind turbines	1
E_{Br} (kWh)—battery capacity	350
η—battery efficiency	0.95
C_{cap}($/kWh)—capital cost of the battery per unit energy	334
A (m^2)—solar panel area	1400
δ—solar panel efficiency	0.2
r^s_{omc}($/kWh)—solar unit cost	0.06
n_g—number of generators	1
R_g (kW)—generator capacity	65
r^g_{omc} ($/kWh)—generator operating cost	0.45
r^{sb} ($/kWh)—reward sold back	0.2

6.2.3 Battery Degradation. The cost of battery degradation stands as another crucial parameter. Batteries represent a significant investment and must be optimally managed to prolong their useful life. As discussed earlier, changes in the storage capacity of batteries occur over time and with different charging/discharging patterns. To assess the reduction in the cost of battery degradation, we adapted the work of Ref. [6] and introduced the concept of battery degradation into their RL-based control policy. Figure 6 illustrates the comparison between the cost of battery degradation and the actual capacity of batteries within both the RL-based control framework and our multiagent control framework. Significant improvements are evident in both comparisons.

Over a span of 100 h, the battery degradation cost amounts to nearly $190 under the RL-based control policy. In contrast, under our proposed MARL-based policy, the degradation cost reduces to $170, representing a notable decrease compared to the RL policy. Likewise, in the RL-based control policy, the actual capacity of the battery diminishes by 0.0006 MWh over the 100-h period. Conversely, under the proposed control policy, the actual capacity of the battery decreases by only 0.0005 MWh during the same time frame.

Figure 7 illustrates the mode variations of the batteries, indicating whether they are charging or discharging. It is evident that the charging/discharging cycles switch more frequently under the RL control policy. In contrast, our updated integrated control takes these variations into account and adjusts the reward accordingly based on the corresponding cycle switches. Agents receive penalties if they contribute to increased battery degradation costs or reduced actual battery capacity. In our integrated system with battery degradation, the cycle switches occur only six times, whereas in the integrated system without considering battery degradation, the cycles switch around 14 times. This difference results in an increase in battery degradation cost and a decrease in actual capacity.

7 Insight From the Trained MARL-Based Policy

RL and DRL are extensively discussed topics across various fields, proving useful in navigating complex environments, especially when the system dynamics are partially or completely unknown. In the realm of renewable energy and microgrids, they have been widely employed to address issues such as demand management, distributed control, and price management. However, a consistent concern exists across domains regarding the nature of DRL algorithms.

To address this concern, we delve into our trained policies and extract the usage pattern. Following training across diverse environments, we examine the energy distribution profile for the manufacturing demand, as shown in Fig. 8. The x-axis represents the time-steps, along which the distribution pattern is observed for 200 time-steps (hours). The y-axis on the left side indicates the amount of energy dispatched by each power source to meet the manufacturing demand, which is represented by the y-axis on the right side.

Based on 1-year data from Solar Energy Local [39] and State Climatologist of Illinois [40], solar power predominantly meets the demand, with wind power supplementing it as needed. Generator power remains largely unused due to two main factors: either the demand is met by solar and wind energy, or the generator's production cost is higher than the utility grid's cheaper rates. While Fig. 8 shows no generator use, it is infrequently utilized, typically only when solar, wind, and battery storage are unavailable, or when grid prices are prohibitively high. Such conditions are rare, explaining the minimal generator use in Fig. 8, where solar, wind, and battery resources are sufficient to meet demand. In the broader dataset, the generator is activated only under these specific conditions—when other energy sources are unavailable, and grid costs are high. Additionally, there are time-steps, such as at 170 and 185, where demand appears to be zero, yet solar or wind power continues to be supplied. This discrepancy results from smoothing the

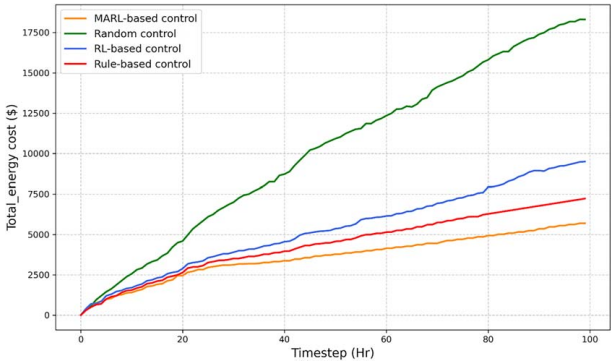

Fig. 4 Total energy consumption under different control policies

demand curve for representation, which shows zero demand while energy is still being consumed.

Overall, the trained policy optimizes control objectives by reducing energy consumption and increasing the throughput through actions taken at both the microgrid and manufacturing levels. However, the rationale behind these actions remains unknown. Therefore, we plot the correlation between the selected continuous actions under the trained policy and the climate conditions, as shown in Fig. 9.

For each climate variable (e.g., solar irradiation, wind speed), daily averages are calculated across 8 days, which is the evaluation period, and plotted alongside the corresponding average continuous actions taken by the agents over a 24-h period. The focus is on continuous actions, specifically the energy supplied by each source to the manufacturing system, the battery, and the grid. The different markers indicate energy sources (solar, wind, and generator) and energy allocation (manufacturing, battery storage, or grid sales). Solar irradiance is normalized for clarity, with the *x*-axis representing a 24-h cycle and the *y*-axis showing solar irradiance and wind speed.

The figure reveals that no energy is supplied by solar, wind, or the generator during the early hours (midnight to 6 a.m.), when the battery or utility grid meets the demand. From 6 a.m. to 6 p.m. (starting at 5 a.m. on the *x*-axis), most energy comes from solar and wind, peaking around noon when solar power fulfills manufacturing demand, charges the battery, and is sold to the grid. After 6 p.m., wind continues to supply some energy, though it is insufficient, prompting the generator to activate as utility prices rise.

While Fig. 9 offers a comprehensive view of energy contributions, the generator markers occasionally overlap, as it supplies excess energy to manufacturing, battery storage, and grid sales simultaneously. However, the general trends remain clear.

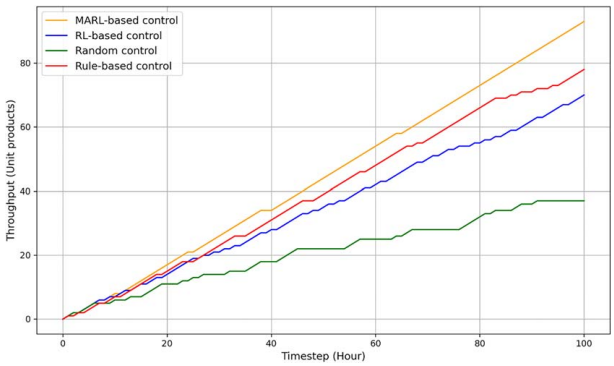

Fig. 5 Comparison of the throughputs based on different control policies

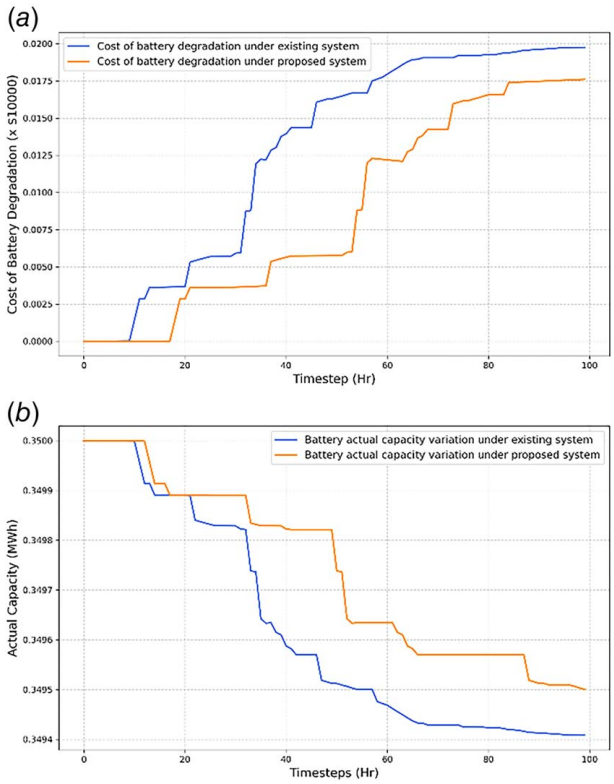

Fig. 6 Comparison under RL- and MARL-based controls of (*a*) battery degradation cost and (*b*) actual capacity

Using the trained MARL policy as a foundation, we propose a rule-based schedule applicable to similar environments with comparable climate conditions and energy needs. The 24-h usage profile is divided into four quarters, as shown in Fig. 10.

From midnight to early morning (12:00 a.m. to 6:00 a.m.), wind energy is the primary source, supplemented by the utility grid or ESS, as generator use is more costly during this period. From morning through early evening (6:00 a.m. to 6:00 p.m.), solar power is prioritized as the optimal energy source, with wind energy used as needed to meet demand. During the evening and into the night (6:00 p.m. to 12:00 a.m.), wind energy remains active, while the generator is engaged only if utility prices exceed generation costs, ensuring uninterrupted operation.

In all scenarios, demand from manufacturing takes first priority, followed by battery storage utilization, with any excess energy sold back to the utility grid. Based on this distribution profile, we

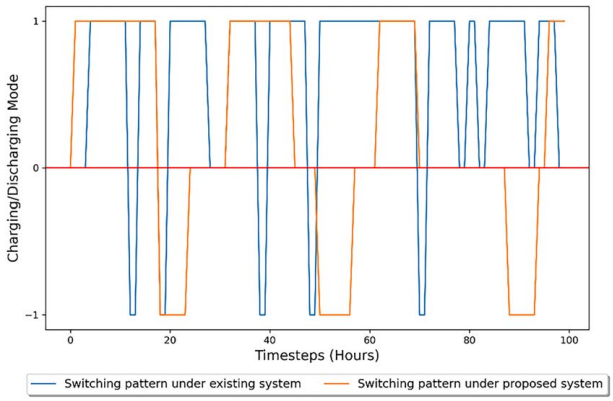

Fig. 7 Battery's charging/discharging patterns

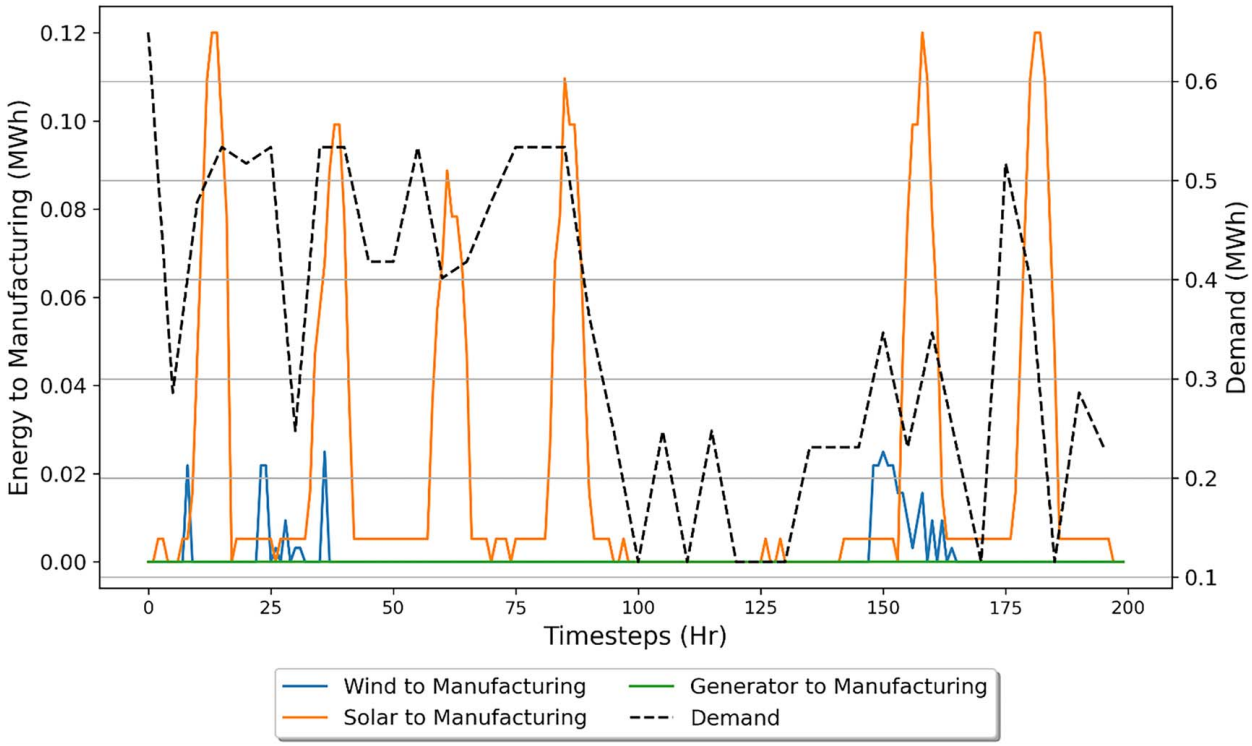

Fig. 8 Energy distribution profile from microgrid to the manufacturing system

establish a rule-based control scheme and evaluate it in our environment, comparing it with RL-based and MARL-based control policies.

This rule-based approach is crafted based on MARL policies for a specific manufacturing system based on Ref. [30] and climatic conditions based on Refs. [39,40]. Consequently, its efficacy might diminish in vastly different settings compared to the current context. While its effectiveness might be limited in very different settings, it provides a strong starting point, especially when computing power is limited. As shown in Fig. 7, this rule-based policy performs just shy of the MARL policy and even outperforms the centralized RL approach in terms of overall energy cost. Therefore, this rule may be applied in similar environments and can be easily implemented in real-world scenarios. It prioritizes cost-effective renewables like solar and wind, with the potential to integrate additional options. However, if other renewable sources become more economical, the rule can be easily adjusted based on their relative costs.

Moreover, while not as adaptable as MARL, the rule-based policy can provide baseline performance without requiring the significant computational overhead associated with training a DRL model. We believe it can serve as a useful tool when transitioning

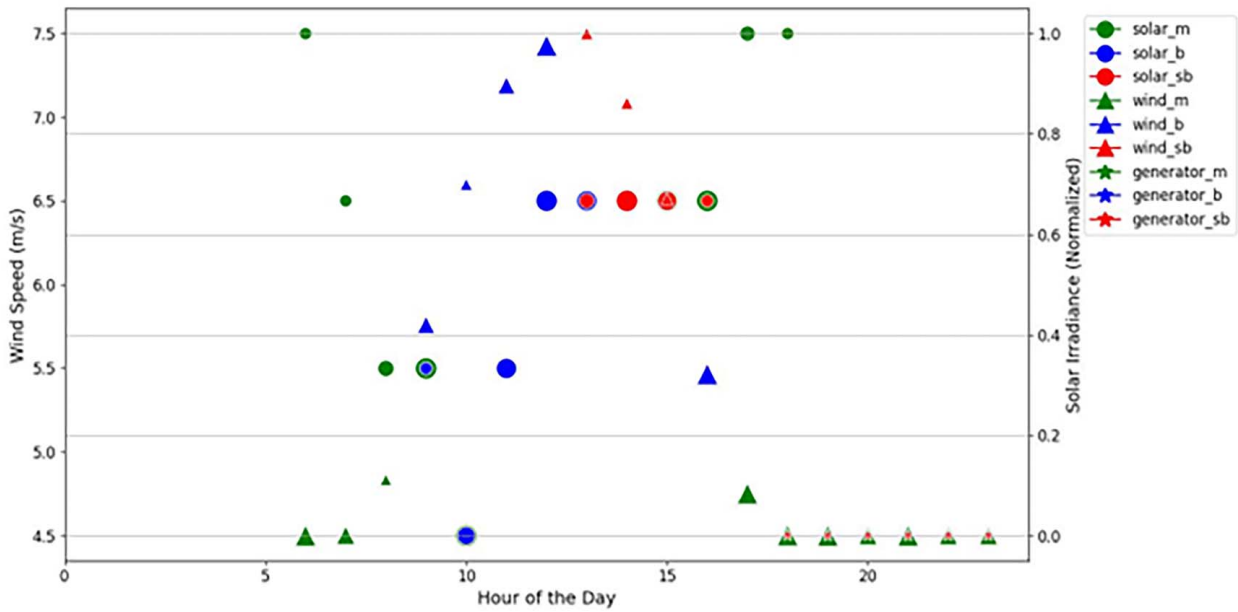

Fig. 9 Correlation among climate conditions and MARL actions

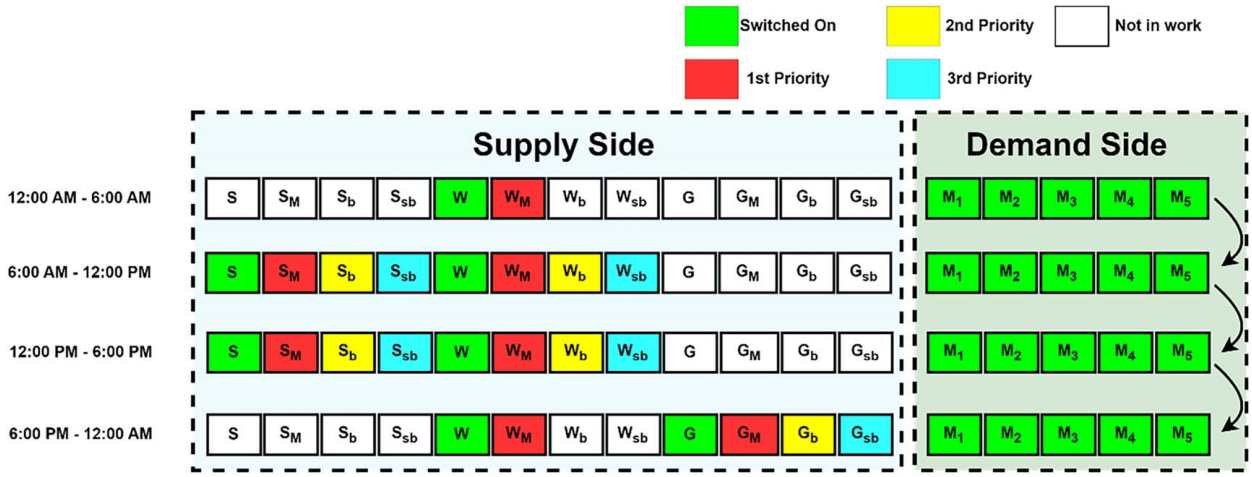

Fig. 10 Average energy usage profile per day

to a more complex MARL policy. Specifically, by using rule-based strategies to guide initial agent behavior, the exploration phase in MARL training can be significantly reduced. Since the agents do not need to explore all possible states from scratch, they can learn more efficiently, which speeds up the convergence of the MARL model. This hybrid approach offers a practical balance between computational simplicity and performance, particularly in the early stages of training, and can be especially useful when experimenting with new environments or applications where training data is limited or costly.

8 Conclusion

This article addresses the integration of a microgrid with a manufacturing system while considering the degradation cost of energy storage systems. The study encompasses several key steps: First, it involves modeling and assessing the dynamics of combined energy consumption, encompassing microgrids, ESS, and manufacturing systems. Subsequently, it situates the supervisory control of the integrated microgrid-manufacturing system within the Dec-POMDP framework, establishing a multiagent system that includes both discrete and continuous agents. To address the control challenge within this framework, the study employs the MADDPG algorithm. Upon completion of the training phase, the policy is evaluated, and a rule-based approach is extracted for straightforward implementation in real-world settings, circumventing the need for complex computational resources. Comparative analysis among our MARL-based approach, centralized RL-based approach, and random control demonstrates that the proposed rule-based control outperforms random and RL-based control strategies.

Funding Data

- The U.S. National Science Foundation (NSF), Grant 2243930.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Pullins, S., 2019, "Why Microgrids Are Becoming an Important Part of the Energy Infrastructure," Electr. J., **32**(5), pp. 17–21.
[2] Uddin, M., Mo, H., Dong, D., Elsawah, S., Zhu, J., and Guerrero, J. M., 2023, "Microgrids: A Review, Outstanding Issues and Future Trends," Energy Strategy Rev., **49**(49), p. 101127.
[3] Mahjoub, S., Chrifi-Alaoui, L., Drid, S., and Derbel, N., 2023, "Control and Implementation of an Energy Management Strategy for a PV–Wind–Battery Microgrid Based on an Intelligent Prediction Algorithm of Energy Production," Energies, **16**(4), p. 1883.
[4] Elmorshedy, M. F., Elkadeem, M. R., Kotb, K. M., Taha, I. B. M., and Mazzeo, D., 2021, "Optimal Design and Energy Management of an Isolated Fully Renewable Energy System Integrating Batteries and Supercapacitors," Energy Convers. Manage., **245**, p. 114584.
[5] Zhou, Q., Shahidehpour, M., Paaso, A., Bahramirad, S., Alabdulwahab, A., and Abusorrah, A., 2020, "Distributed Control and Communication Strategies in Networked Microgrids," IEEE Commun. Surv. Tutor., **22**(4), pp. 2586–2633.
[6] Yang, J., Sun, Z., Hu, W., and Steinmeister, L., 2022, "Joint Control of Manufacturing and Onsite Microgrid System via Novel Neural-Network Integrated Reinforcement Learning Algorithms," Appl. Energy, **315**, p. 118982.
[7] Huang, M., Lin, X., Feng, Z., Wu, D., and Shi, Z., 2023, "A Multi-Agent Decision Approach for Optimal Energy Allocation in Microgrid System," Electr. Power Syst. Res., **221**, p. 109399.
[8] Neri, A., Butturi, M. A., Lolli, F., and Gamberini, R., 2023, "Inter-Firm Exchanges, Distributed Renewable Energy Generation, and Battery Energy Storage System Integration via Microgrids for Energy Symbiosis," J. Clean. Prod., **414**, p. 137529.
[9] Dashtdar, M., Bajaj, M., and Hosseinimoghadam, S. M. S., 2022, "Design of Optimal Energy Management System in a Residential Microgrid Based on Smart Control," Smart Sci., **10**(1), pp. 25–39.
[10] Ahmad, T., and Zhang, D., 2020, "A Critical Review of Comparative Global Historical Energy Consumption and Future Demand: The Story Told So Far," Energy Rep., **6**, pp. 1973–1991.
[11] Islam, M. M., and Yao, X., 2016, "Simulation-Based Investigation for the Application of Microgrid With Renewable Sources in Manufacturing Systems Towards Sustainability," Proceedings of the International Annual Conference of the American Society for Engineering Management, Charlotte, NC, Oct. 26.
[12] Zhong, X., Islam, M., Xiong, H., and Sun, Z., 2017, "Design the Capacity of Onsite Generation System With Renewable Sources for Manufacturing Plant," Proc. Comput. Sci., **114**, pp. 433–440.
[13] Yao, T., Jiang, D., Xin, R., Wu, J., and Sun, S., 2020, "Load Prediction of Microgrid Optimal Operation Based on Improved Algorithm in Machine Learning," Int. J. Mechatron. Appl. Mech., **7**(18), pp. 124–128.
[14] Faraji, J., Ketabi, A., Hashemi-Dezaki, H., Shafie-Khah, M., and Catalao, J. P., 2020, "Optimal Day-Ahead Self-Scheduling and Operation of Prosumer Microgrids Using Hybrid Machine Learning-Based Weather and Load Forecasting," IEEE Access, **8**, pp. 157284–157305.
[15] Luo, X., and Mahdjoubi, L., 2024, "Towards a Blockchain and Machine Learning-Based Framework for Decentralised Energy Management," Energy Build., **303**, p. 113757.
[16] Rosero, D. G., Díaz, N. L., and Trujillo, C. L., 2021, "Cloud and Machine Learning Experiments Applied to the Energy Management in a Microgrid Cluster," Appl. Energy, **304**, p. 117770.
[17] Rosero, D. G., Sanabria, E., Díaz, N. L., Trujillo, C. L., Luna, A., and Andrade, F., 2023, "Full-Deployed Energy Management System Tested in a Microgrid Cluster," Appl. Energy, **334**, p. 120674.
[18] Mobtahej, M., Barzegaran, M., and Esapour, K., 2023, "A Novel Three-Stage Demand Side Management Framework for Stochastic Energy Scheduling of Renewable Microgrids," Sol. Energy, **256**, pp. 32–43.

[19] Marimuthu, R., 2023, "Review on Advanced Control Techniques for Microgrids," Energy Rep., **10**, pp. 3054–3072.
[20] Azarhooshang, A., Sedighizadeh, D., and Sedighizadeh, M., 2021, "Two-Stage Stochastic Operation Considering Day-Ahead and Real-Time Scheduling of Microgrids With High Renewable Energy Sources and Electric Vehicles Based on Multi-Layer Energy Management System," Electr. Power Syst. Res., **201**, p. 107527.
[21] Chung, C.-H., Jangra, S., Lai, Q., and Lin, X., 2020, "Optimization of Electric Vehicle Charging for Battery Maintenance and Degradation Management," IEEE Trans. Transp. Electrif., **6**(3), pp. 958–969.
[22] Silva, J. A. A., López, J. C., Guzman, C. P., Arias, N. B., Rider, M. J., and da Silva, L. C. P., 2023, "An IoT-Based Energy Management System for AC Microgrids With Grid and Security Constraints," Appl. Energy, **337**, p. 120904.
[23] Miletić, M., Pandžić, H., and Yang, D., 2020, "Operating and Investment Models for Energy Storage Systems," Energies, **13**(18), p. 4600.
[24] Dinata, N. F. P., Ramli, M. A. M., Jambak, M. I., Sidik, M. A. B., and Alqahtani, M. M., 2024, "Designing an Optimal Microgrid Control System Using Deep Reinforcement Learning: A Systematic Review," Eng. Sci. Technol. Int. J., **51**, p. 101651.
[25] Lu, R., Hong, S. H., and Zhang, X., 2018, "A Dynamic Pricing Demand Response Algorithm for Smart Grid: Reinforcement Learning Approach," Appl. Energy, **220**, pp. 220–230.
[26] Kofinas, P., Vouros, G., and Dounis, A. I., 2018, "Energy Management in Solar Microgrid via Reinforcement Learning Using Fuzzy Reward," Adv. Build. Energy Res., **12**(1), pp. 97–115.
[27] Aaltonen, H., Sierla, S., Subramanya, R., and Vyatkin, V., 2021, "A Simulation Environment for Training a Reinforcement Learning Agent Trading a Battery Storage," Energies, **14**(17), p. 5587.
[28] Li, Z., Wu, L., Xu, Y., Moazeni, S., and Tang, Z., 2021, "Multi-Stage Real-Time Operation of a Multi-Energy Microgrid With Electrical and Thermal Energy Storage Assets: A Data-Driven MPC-ADP Approach," IEEE Trans. Smart Grid, **13**(1), pp. 213–226.
[29] Waseem, M., and Chang, Q., 2023, "Adaptive Mobile Robot Scheduling in Multiproduct Flexible Manufacturing Systems Using Reinforcement Learning," ASME J. Manuf. Sci. Eng., **145**(12), p. 121005.
[30] Waseem, M., and Chang, Q., 2024, "From Nash Q-Learning to Nash-MADDPG: Advancements in Multiagent Control for Multiproduct Flexible Manufacturing Systems," J. Manuf. Syst., **74**, pp. 129–140.
[31] Xiong, J., Wang, Q., Yang, Z., Sun, P., Han, L., Zheng, Y., Fu, H., Zhang, T., Liu, J., and Liu, H., 2018, "Parametrized Deep q-Networks Learning: Reinforcement Learning With Discrete-Continuous Hybrid Action Space," preprint arXiv:1810.06394.
[32] Fu, H., Tang, H., Hao, J., Lei, Z., Chen, Y., and Fan, C., 2019, "Deep Multi-Agent Reinforcement Learning With Discrete-Continuous Hybrid Action Spaces," preprint arXiv:1903.04959.
[33] Hua, H., Zhao, R., Wen, G., and Wu, K., 2023, "A Further Exploration of Deep Multi-Agent Reinforcement Learning With Hybrid Action Space," International Conference on Artificial Neural Networks, Belgrade, Serbia, Sept. 22–24, Springer.
[34] Fan, Z., Su, R., Zhang, W., and Yu, Y., 2019, "Hybrid Actor-Critic Reinforcement Learning in Parameterized Action Space," preprint arXiv:1903.01344.
[35] Huang, C., Zhang, H., Wang, L., Luo, X., and Song, Y., 2022, "Mixed Deep Reinforcement Learning Considering Discrete-Continuous Hybrid Action Space for Smart Home Energy Management," J. Modern Power Syst. Clean Energy, **10**(3), pp. 743–754.
[36] Neunert, M., Abdolmaleki, A., Wulfmeier, M., Lampe, T., Springenberg, T., Hafner, R., Romano, F., Buchli, J., Heess, N., and Riedmiller, M., 2020, "Continuous-Discrete Reinforcement Learning for Hybrid Control in Robotics," Conference on Robot Learning, Osaka, Japan, Oct. 30, PMLR.
[37] Zou, J., Chang, Q., Arinez, J., and Xiao, G., 2017, "Data-Driven Modeling and Real-Time Distributed Control for Energy Efficient Manufacturing Systems," Energy, **127**, pp. 247–257.
[38] Wen, X., Fu, Y., Yang, W., Wang, H., Zhang, Y., and Sun, C., 2023, "An Effective Hybrid Algorithm for Joint Scheduling of Machines and AGVs in Flexible Job Shop," Meas. Control, **56**(9–10), p. 00202940231173750.
[39] Local, S. E., 2016, "Solar Energy Data and Resources in the US." https://solarenergylocal.com/.
[40] Angel, J., 2000, "State Climatologist Office for Illinois." http://www.isws.illinois.edu/atmos/statecli/wind/wind.htm.
[41] Appelman, M., Venugopal, P., and Rietveld, G., 2022, "Impact of Discharge Current Profiles on Li-Ion Battery Pack Degradation," 2022 IEEE 20th International Power Electronics and Motion Control Conference (PEMC), Brasov, Romania, Sept. 25–29.
[42] Mughees, N., Jaffery, M. H., Mughees, A., Ansari, E. A., and Mughees, A., 2023, "Reinforcement Learning-Based Composite Differential Evolution for Integrated Demand Response Scheme in Industrial Microgrids," Appl. Energy, **342**, p. 121150.
[43] Zhou, C., Qian, K., Allan, M., and Zhou, W., 2011, "Modeling of the Cost of EV Battery Wear Due to V2G Application in Power Systems," IEEE Trans. Energy Convers., **26**(4), pp. 1041–1050.
[44] Koller, M., Borsche, T., Ulbig, A., and Andersson, G., 2013, "Defining a Degradation Cost Function for Optimal Control of a Battery Energy Storage System," 2013 IEEE Grenoble Conference, Grenoble, France, June 16–20.
[45] Li, J., and Meerkov, S. M., 2008, *Production Systems Engineering*, Springer Science & Business Media, New York.
[46] Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I., 2017, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," Adv. Neural Inf. Process. Syst., **30**, pp. 6382–6393.