energies

MDPI

Article

Deep Reinforcement Learning for Battery Energy Storage Optimization and Residential Decarbonization in Grid-Deficient Environments: An Iraqi Case Study

Ahmed Mohammed [1], Badr M. Abdullah [2], Ali Shubbar [2,*], Qian Zhang [1], Omar Aldhaibani [3], Jeff Cullen [2] and Amer Salih [1]

[1] School of Engineering, Liverpool John Moores University, Liverpool L3 3AF, UK; a.s.mohammed@ljmu.ac.uk (A.M.); a.m.salih@ljmu.ac.uk (A.S.)
[2] School of Civil Engineering and Built Environment, Liverpool John Moores University, Liverpool L3 3AF, UK; b.m.abdullah@ljmu.ac.uk (B.M.A.)
[3] School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool L3 3AF, UK
* Correspondence: a.a.shubbar@ljmu.ac.uk

Abstract

In grid-deficient environments, residential energy systems face severe carbon emission penalties due to mandatory reliance on diesel standby generators during supply interruptions. In Iraq, summer peak loads routinely exceed grid capacity, triggering prolonged generator operation and dramatically increasing household carbon footprints. This study presents a deep Q-network (DQN) reinforcement learning framework for intelligent battery energy storage system (BESS) scheduling, targeting carbon emissions reduction through strategic peak shaving. The DQN agent learns optimal battery dispatch strategies by internalizing diurnal patterns in load and solar generation through temporal state features, enabling anticipatory control without requiring explicit external forecasting models. The system is trained on one-year operational data from a representative Iraqi residential installation and evaluated over the critical summer period (122 days, 35.5% grid unavailability). The results demonstrate a 54.8% CO_2 reduction (306.5 kg versus 677.4 kg baseline), a 25.5% reduction in generator runtime, and a 23.7% reduction in operating costs for the studied configuration. The learned policy approaches 89.6% of perfect-foresight MILP performance while executing 35,000 times faster. A reward function sensitivity analysis across five weighting schemes confirms that the 20:1 carbon-to-cost priority ratio optimally balances environmental and economic objectives. Ablation studies quantify the mechanism contributions: anticipatory pre-charging accounts for 58% of the total improvement, discharge optimization for 44%, and real-time PV coordination for 22%. These findings establish DQN-based BESS optimization as a practically deployable decarbonization approach for residential systems in grid-constrained developing regions.

Keywords: deep reinforcement learning; deep Q-network; battery energy storage; carbon emissions reduction; hybrid renewable energy; grid instability; peak shaving

Academic Editor: Helena M. Ramos

Received: 10 January 2026
Revised: 19 February 2026
Accepted: 25 February 2026
Published: 1 March 2026

1. Introduction

The global transition toward sustainable energy systems has been intensifying the need for intelligent residential energy management strategies that can effectively balance energy supply and demand while minimizing environmental impacts. Residential buildings account for approximately 30% of the global electricity consumption, and this proportion

continues to rise with increasing electrification of heating, cooling, and transportation systems [1]. While the transmission and distribution infrastructure faces capacity limitations in grid-constrained regions, the challenge of managing residential energy demand becomes particularly acute. Combined with the intermittent nature of distributed renewable energy resources, such as rooftop photovoltaic (PV) systems, sophisticated control strategies are necessary that can optimize energy flows in real time while maintaining grid stability and reducing carbon emissions [2,3].

Home Energy Management Systems (HEMSs) have emerged as a critical enabling technology for addressing these challenges by coordinating the operation of distributed energy resources (DERs), battery energy storage systems (BESSs), and flexible loads within residential environments [4]. The fundamental objective of an HEMS is to minimize electricity costs, enhance the self-consumption of locally generated renewable energy, reduce peak demand on the grid, and improve overall energy efficiency while maintaining occupant comfort [5]. However, the design and implementation of effective HEMSs face several interconnected challenges: the stochastic nature of renewable energy generation, the unpredictability of residential load patterns, the complexity of multi-objective optimization under uncertainty, and the computational burden of real-time decision-making [6]. These challenges are further compounded in grid-constrained regions, where the ability to import electricity from the grid may be limited during peak demand periods, necessitating more sophisticated local energy management strategies [7].

Early approaches to residential energy management relied predominantly on rule-based control strategies, which employ predefined heuristics and threshold-based decision rules to schedule appliances and manage battery charging and discharging [8]. While rule-based methods offer simplicity and interpretability, they suffer from fundamental limitations that restrict their effectiveness in dynamic and uncertain environments. Rule-based controllers typically operate in a reactive manner, responding to current system states without anticipating future conditions or learning from historical patterns [9]. This reactive nature leads to suboptimal decisions in scenarios involving time-varying electricity prices, fluctuating renewable generation, and uncertain load profiles [10]. Furthermore, rule-based approaches lack the flexibility to adapt to changing environmental conditions, seasonal variations, or evolving occupant behavior patterns, requiring manual recalibration and expert knowledge for each specific deployment context [2].

To address the limitations of rule-based methods, optimization-based approaches have been extensively investigated in the literature. Model predictive control (MPC), mixed-integer linear programming (MILP), and dynamic programming techniques have been applied to formulate HEMSs as constrained optimization problems, seeking to minimize cost functions subject to physical and operational constraints [2,8,11]. Compressive receding horizon approaches exploit week-ahead PV and weather forecasts for improving self-consumption and reduced grid stress, while stochastic online forecast-and-optimize frameworks integrate uncertainty estimation for robust real-time dispatch in virtual power plants [11]. However, optimization-based approaches face significant practical challenges. First, they require precise mathematical models of building thermal dynamics, appliance behavior, and battery degradation characteristics, which are often difficult to obtain or maintain in real-world deployments. Recent reviews have categorized stochastic, robust, and fuzzy optimization techniques for managing uncertainties in microgrids, highlighting the trade-offs between computational complexity and solution quality [3]. Second, the computational complexity of solving large-scale optimization problems in real time is prohibitive, particularly when considering the combinatorial nature of appliance scheduling and the nonlinear dynamics of battery systems [1]. Third, optimization-based methods exhibit limited generalization capabilities as they struggle to handle the nonlinear behavior

of residential systems, the volatility of renewable energy sources, and the heterogeneity of end-user preferences [5].

Machine learning and artificial intelligence techniques have gained significant traction in residential energy management research over the past decade, offering the potential to overcome many limitations of conventional approaches. Recent comprehensive reviews have surveyed computational intelligence approaches for microgrid HEMSs, highlighting challenges in scalability, privacy, and real-world deployment [6], as well as AI/DRL integration in smart microgrids with an emphasis on testbeds, decentralization, and cybersecurity [12]. Among various machine learning paradigms, supervised learning methods—including Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM) networks, and ensemble methods such as Random Forest—have been widely applied to load forecasting and demand prediction tasks [3]. Accurate load forecasting is a critical prerequisite for effective energy management as it enables proactive scheduling decisions and reduces the uncertainty inherent in residential energy systems [13,14]. Advanced architectures combining non-intrusive load monitoring (NILM) with multi-objective scheduling have demonstrated significant operational and environmental benefits [13], while climate-adaptive frameworks have integrated CNN–BiLSTM forecasting with multi-objective optimization to address temperature-sensitive loads [14]. Recent studies have demonstrated that deep learning architectures can achieve superior prediction accuracy compared to traditional statistical methods [1]. However, most existing approaches treat forecasting and control as separate sequential tasks rather than as integrated components of a unified decision-making framework [3,6].

Reinforcement learning (RL) has emerged as a particularly promising paradigm for HEMS control as it enables autonomous agents to learn optimal decision-making policies through interaction with the environment without requiring explicit system models [5]. Deep reinforcement learning (DRL) addresses the scalability limitations of classical RL by leveraging deep neural networks as function approximators, enabling learning in continuous and high-dimensional state spaces [1,4]. The deep Q-network (DQN) algorithm, which combines Q-learning with deep neural networks and experience replay mechanisms, has been successfully applied to various HEMS control tasks [5,10,15]. The applications include grid-tied PV-battery microgrids for minimizing transaction costs and battery degradation and energy purchase optimization integrating renewable forecasts with market prices [15]. Comparative studies have evaluated different DRL algorithms (DQN, DDPG, TD3, PPO, and SAC) for HEMS applications, demonstrating trade-offs between sample efficiency, convergence speed, and policy optimality [16,17]. Multi-agent DRL frameworks using DQN for distributed energy management and demand response have achieved reductions in average daily bills ranging from 22.1% to 48.8% for different prosumer profiles while increasing service provider profit by 25.7% and reducing reserve power consumption by 16% [5]. Advanced DQN variants, including Dueling DQN, have been explored to improve learning stability and sample efficiency, achieving 5.6% energy cost savings compared to rule-based methods and demonstrating convergence times that are 71% faster than standard DQN [10].

Multi-agent reinforcement learning (MARL) frameworks have been investigated to address the distributed nature of residential energy systems and enable coordination among multiple households or prosumers [5,18]. Scalable MARL approaches for distributed control of residential energy flexibility have demonstrated value creation through reductions in energy import costs, network congestion, battery depreciation, and greenhouse gas emissions [18]. Transactive energy frameworks incorporating storage degradation in bidding models have demonstrated cost savings while reducing battery wear for residential DER aggregation, while decentralized MARL approaches for incentive-based DR have

shown potential for preserving privacy in aggregator-managed cohorts [5]. Multiagent RL community aggregators mitigate peak rebounds from price-based DR through decomposition and renewable forecasting [19], while actor–critic DRL with federated learning scales DR to distribution feeders [20]. Real-time autonomous DR management using TD3 has been validated in real-world case studies [21]. Recent advances integrating evolutionary game theory with DRL provide theoretical foundations for extending single-agent frameworks toward multi-agent strategic coordination under market uncertainty [22].

Battery energy storage systems (BESSs) play a pivotal role in residential energy management by decoupling energy generation from consumption, enabling load shifting, peak shaving, and enhanced self-consumption of renewable energy [1,8]. The optimization of BESS operation involves complex trade-offs among multiple objectives: minimizing electricity costs, maximizing PV self-consumption, reducing battery degradation, and maintaining sufficient reserve capacity for backup power [4,8]. Recent reviews emphasize the critical importance of incorporating battery aging models and techno-economic assessments into BESS optimization frameworks to ensure long-term viability and profitability [23]. However, most existing approaches optimize battery operation based solely on economic objectives, neglecting the environmental dimension of carbon emissions reduction [2]. Furthermore, battery degradation models are often oversimplified or entirely omitted in RL formulations, potentially leading to control policies that accelerate battery aging [8,23].

The integration of carbon emissions awareness into HEMS represents a critical yet underexplored research direction. While several review papers acknowledge the potential for HEMSs to reduce carbon emissions through demand response and renewable energy integration, few studies explicitly incorporate carbon intensity signals into their control objectives or reward functions. The carbon intensity of grid electricity varies significantly over time, depending on the generation mix and the dispatch of fossil fuel versus renewable power plants. By shifting flexible loads to periods of low carbon intensity and prioritizing self-consumption of zero-carbon PV generation, HEMSs can achieve substantial emissions reductions beyond those achievable through cost minimization alone [2].

Peak shaving—reducing maximum power demand during peak periods—is a critical function of HEMSs in grid-constrained regions, where distribution transformers and transmission lines may be capacity-limited [4,9]. Excessive peak demand can lead to grid instability, voltage fluctuations, equipment overloading, and increased electricity costs due to demand charges or time-of-use pricing structures [9]. Reinforcement learning-based peak-shaving strategies have been investigated, with studies demonstrating effective reductions in peak load while maintaining occupant comfort [9,19]. However, most peak-shaving studies focus on single-objective optimization and do not consider the integration of load forecasting, battery optimization, and carbon emissions reduction within a unified framework [10].

Despite the substantial progress in machine learning and reinforcement learning for residential energy management, several critical research gaps remain. First, most existing studies treat load forecasting and control as separate sequential tasks rather than integrated components of a unified decision-making framework [6]. This separation can lead to suboptimal performance as forecasting models are optimized for prediction accuracy rather than for their impact on control decisions [3]. Second, the majority of DRL-based HEMS studies focus on single-objective optimization (typically cost minimization) and do not explicitly incorporate carbon emissions reduction into the reward structure [5]. Third, many studies rely on simplified battery models that neglect degradation dynamics, potentially leading to control policies that accelerate battery aging [8,23]. Fourth, the reactive nature of many existing approaches limits their ability to anticipate future conditions and make proactive scheduling decisions [4]. Fifth, insufficient attention has been paid to appliance-level en-

ergy modeling and disaggregation, which are essential for granular control and targeted demand response [13]. Adaptive reinforcement learning techniques have been applied to dynamic appliance scheduling, demonstrating potential for real-time optimization of flexible loads [24].

This paper proposes a DQN reinforcement learning framework for intelligent battery dispatch in grid-deficient residential environments, addressing the identified research gaps through the following contributions. First, a multi-objective reward function explicitly prioritizes carbon emissions reduction (20:1 carbon-to-cost weighting), enabling environmental optimization beyond conventional cost-minimization approaches. Second, temporal state features (hour-of-day, day-of-year, and target SOC trajectory) allow the agent to internalize anticipatory control patterns without requiring explicit external forecasting models, reducing system complexity and data dependencies in resource-constrained deployment contexts. Third, the framework is demonstrated under extreme grid instability (35.5% unavailability or prolonged multi-hour outages) that is substantially more severe than the conditions reported in the prior HEMS literature, with validation across four seasonal scenarios. Fourth, quantitative mechanism decomposition via ablation studies and reward sensitivity analysis rigorously validates the relative contribution of each operational strategy to total emissions reduction. Fifth, a techno-economic sizing analysis identifies 10 kWh as the optimal battery capacity for the studied configuration, with a levelized cost of emissions reduction (LCER) of \$38/ton CO_2. These findings are illustrative of a specific Iraqi residential configuration and should be generalized with consideration of local system sizing, outage patterns, and load characteristics. Empirical evidence from large-scale residential demonstrations has validated the feasibility of zero-carbon districts with PV and micro-cogeneration, providing insights into occupant engagement and operational performance [25].

2. Materials and Methods

The methodology presented in this study addressed the optimal dispatch of battery energy storage within a hybrid renewable energy system subjected to severe grid instability. The research was conducted through simulation-based experimentation, wherein a deep Q-network (DQN) reinforcement learning agent was trained to minimize carbon dioxide emissions while maintaining energy reliability for a typical residential load in Iraq. This section provides a comprehensive description of the system architecture, component modeling procedures, the DQN-based control framework, emissions quantification methodology, and simulation environment configuration. All the methodological choices were made to ensure reproducibility and alignment with practical operational constraints encountered in grid-deficient environments.

2.1. System Architecture and Component Modeling

The hybrid energy system architecture was designed to represent a realistic residential installation in Iraq, where prolonged grid outages during summer months necessitate backup power generation and strategic energy storage management. The system comprised four interconnected subsystems: a solar photovoltaic (PV) generation array, a lithium-ion battery energy storage system (BESS), a diesel-fueled standby generator, and a connection to an unreliable electrical grid. The integration of these components was designed to maximize renewable energy utilization while ensuring continuous power supply despite frequent and unpredictable grid failures. Figure 1 illustrates the complete system topology, showing the power flow pathways, control interfaces, and the hierarchical dispatch logic that prioritized renewable energy utilization while minimizing fossil fuel dependency. The figure depicts the PV array and bidirectional battery to enable both charging and

discharging operations; both are connected to a smart inverter, the diesel generator as a backup AC source with automatic transfer switching capability and the grid connection point with a binary availability indicator representing the stochastic nature of grid outages. The residential load connection point gets power from any combination of the four sources. The DQN agent is the main controller that receives state inputs (battery state of charge, load demand, PV generation, grid availability status, and time-of-day information) and sends out battery power commands. The arrows in the diagram show that the battery can send and receive power in both directions, while the PV generation, grid import, and generator output can only send power in one direction. Control signal paths are separate from power flow paths.

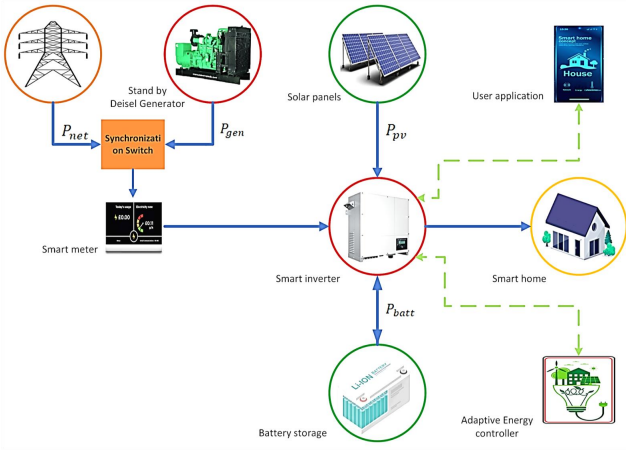

Figure 1. Hybrid energy system architecture.

The modeling approach adopted a quasi-steady-state assumption, wherein power balance was enforced at each 15 min time step without consideration of transient dynamics or inverter switching behavior. Sub-minute transient phenomena-inverter switching, automatic transfer switch (ATS) response (10–300 ms), and power ramping (5–10 kW/s) occur four or more orders of magnitude faster than the 15 min decision horizon and are appropriately managed by lower-level inverter control loops; they are therefore intentionally excluded from the energy management model. Bidirectional inverter efficiency varies between 94 and 97% across the operating range; the adopted constant value of $\eta_{\mathrm{ch}} = \eta_{\mathrm{dch}} = 0.95$ represents a conservative average introducing a maximum 3% error in energy accounting without affecting the directional correctness of the optimization. This temporal resolution was selected to balance computational tractability with sufficient granularity to capture diurnal solar and load variations. The quasi-steady-state assumption is justified for energy management optimization purposes while acknowledging that high-fidelity deployment would require integration with lower-level inverter control loops managing voltage and frequency regulation.

2.2. Photovoltaic Generation System

The photovoltaic generation subsystem was represented by empirical production data rather than first-principles irradiance-to-power conversion models. A grid-connected rooftop PV array with a nominal peak capacity of 8.0 kWp was assumed, typical of residential installations capable of offsetting a significant portion of daytime consumption. The PV output power $P_{\mathrm{pv}}(t)$ was obtained from measured historical generation profiles recorded at a site in Iraq with similar climatic conditions, spanning a complete calendar year to capture seasonal irradiance variations. The dataset included 35,040 fifteen-minute interval measurements providing comprehensive coverage of daily, weekly, and seasonal generation patterns.

The choice to use empirical generation data was motivated by two considerations. First, the measured data inherently captured site-specific shading, soiling, and atmospheric effects that are difficult to model accurately from first principles. Second, the research focus was on developing an optimal dispatch strategy that could operate with realistic generation forecasts, making actual measured profiles more representative than idealized model outputs. No curtailment of PV generation was permitted except when battery storage reached maximum capacity and instantaneous load demand was fully satisfied by solar production, reflecting the priority given to renewable energy utilization.

2.2.1. Battery Energy Storage System

The battery energy storage system was modeled as a lithium-ion battery bank with a nominal capacity of 10 kWh and a maximum continuous power rating of 5 kW for both charging and discharging operations. The battery state of charge (SOC) dynamics were governed by a discrete-time difference equation that accounted for charging efficiency, discharging efficiency, and self-discharge losses:

$$\mathrm{SOC}(t+1) = \mathrm{SOC}(t) \cdot (1 - \sigma \Delta t) + \frac{P_{\mathrm{batt}}(t) \cdot \Delta t}{E_{\mathrm{nom}}} \cdot \begin{cases} \eta_{\mathrm{ch}} & \text{if } P_{\mathrm{batt}}(t) > 0 \\ \frac{1}{\eta_{\mathrm{dch}}} & \text{if } P_{\mathrm{batt}}(t) < 0 \end{cases} \quad (1)$$

where $\mathrm{SOC}(t)$ represents the state of charge at time step t (dimensionless, ranging from 0 to 1), σ denotes the self-discharge rate (0.0001 h^{-1}), Δt is the simulation time step (0.25 h), $P_{\mathrm{batt}}(t)$ is the battery power (kW, positive for charging and negative for discharging), E_{nom} is the nominal battery capacity (10 kWh), η_{ch} is the charging efficiency (0.95), and η_{dch} is the discharging efficiency (0.95).

The SOC was constrained to remain within operational limits to prevent battery degradation:

$$\mathrm{SOC}_{\mathrm{min}} \leq \mathrm{SOC}(t) \leq \mathrm{SOC}_{\mathrm{max}} \quad (2)$$

where $\mathrm{SOC}_{\mathrm{min}} = 0.2$ and $\mathrm{SOC}_{\mathrm{max}} = 0.9$ were enforced as hard constraints. These limits represented a practical operating window that balanced usable capacity (70% of nominal) against cycle life preservation). The battery power command was further constrained by the maximum continuous power rating:

$$-P_{\mathrm{max}} \leq P_{\mathrm{batt}}(t) \leq P_{\mathrm{max}} \quad (3)$$

where $P_{\mathrm{max}} = 5$ kW. When the desired battery power computed by the DQN policy would violate SOC or power limits, the command was clipped to the nearest feasible value, and the actual executed power was used for subsequent state transitions and reward calculations.

The marginal degradation cost of \$0.05/kWh throughput was derived from lifecycle cost amortization of a 10 kWh LFP system (\$5000 installed) over 6000 equivalent full cycles to 80% end-of-life capacity. Each 1% capacity loss reduces system value by approximately \$50, and approximately 1000 kWh cumulative throughput produces 1% capacity fade under typical cycling conditions, yielding \$0.05/kWh. This linear approximation neglects depth-of-discharge (DoD) nonlinearities; the DQN policy's implicitly shallower cycling patterns (mean DoD 0.31 versus 0.42 for the baseline, as characterized in Section 3.3) partially mitigate the throughput increase and its degradation implications.

2.2.2. Diesel Generator

The diesel generator was modeled as a dispatchable backup power source with a rated capacity of 6 kW, capable of supplying the residential load during simultaneous grid outages and insufficient renewable generation. The generator was assumed to operate in

an on–off mode rather than continuous modulation. When activated, the generator output power $P_{gen}(t)$ was set equal to the net load deficit:

$$P_{gen}(t) = \max(0, P_{net}(t)) \quad (4)$$

where $P_{net}(t) = L(t) - P_{pv}(t) + P_{batt}(t)$ represented the net power demand after accounting for PV generation and battery contribution.

The fuel consumption rate was modeled using a piecewise-linear approximation:

$$F(t) = \begin{cases} 0 & \text{if } P_{gen}(t) = 0 \\ a + b \cdot P_{gen}(t) & \text{if } P_{gen}(t) > 0 \end{cases} \quad (5)$$

where $F(t)$ is the fuel consumption rate (liters per hour), $a = 0.8$ L/h represents the no-load fuel consumption, and $b = 0.25$ L/kWh is the incremental fuel consumption coefficient. The operational cost was calculated by multiplying fuel consumption by the local diesel fuel price ($0.50 per liter), with an additional startup cost penalty of $0.10 per start to account for mechanical wear.

2.2.3. Grid Connection and Outage Modeling

The electrical grid connection was modeled as a binary availability variable $G(t) \in \{0, 1\}$, where $G(t) = 1$ indicated grid availability and $G(t) = 0$ indicated an outage. The grid availability time series was derived from historical outage records collected in Iraq, capturing the stochastic and often prolonged nature of grid failures during peak summer demand periods. When the grid was available, power could be imported at a fixed electricity tariff of $0.08 per kWh. No export of excess generation to the grid was permitted, reflecting the absence of net metering policies in the study region.

When the grid was unavailable, the system operated in islanded mode, relying exclusively on PV generation, battery discharge, and diesel generator output to meet load demand. The grid outage patterns exhibited strong temporal correlation, with outages often persisting for multiple consecutive hours during peak afternoon demand periods, creating a challenging control problem for the DQN agent.

2.3. Load Demand Data and Preprocessing

The residential load demand profile $L(t)$ was constructed from measured consumption data representative of a typical Iraqi household with air conditioning, refrigeration, lighting, and electronic appliances. The annual load profile exhibited strong diurnal patterns with morning and evening peaks and a pronounced mid-day peak during summer months due to air conditioning load. The average daily consumption was approximately 25 kWh, with peak instantaneous demand reaching 4–5 kW. The load data were preprocessed to remove anomalies and synchronized with the PV generation and grid availability time series at 15 min resolution. No load shedding was permitted; the system was required to meet 100% of load demand at all times.

2.4. Dataset Partitioning

The complete one-year dataset (35,040 samples, 365 days at 15 min resolution) was partitioned into training and test subsets. The summer period (May through August, 122 days, 11,713 samples) was reserved exclusively for testing to evaluate performance under the most challenging operational conditions with highest grid outage frequency (35.5% unavailability). The remaining 243 days (23,327 samples) comprising spring, autumn, and winter periods were used for training. This partitioning strategy ensured that the DQN agent was tested on out-of-sample data representing the critical high-emissions season.

2.5. Deep Q-Network Control Framework

The battery energy storage dispatch problem was formulated as a Markov Decision Process (MDP) and solved using the deep Q-network (DQN) reinforcement learning algorithm. The DQN framework enabled the agent to learn an optimal policy through trial-and-error interaction with the simulated environment without requiring explicit knowledge of system dynamics or future conditions.

2.5.1. State Space Definition

The state space $\mathcal{S}$ was designed to provide the DQN agent with sufficient information to make informed dispatch decisions. Each state vector $s_t \in \mathbb{R}^7$ at time step t comprised seven continuous features:

1. Battery State of Charge : $\mathrm{SOC}(t) \in [0.2, 0.9]$.
2. Normalized Load Demand: $L_{\mathrm{norm}}(t) = L(t)/L_{\max}$, where $L_{\max} = 6$ kW.
3. Normalized PV Generation: $P_{\mathrm{pv,norm}}(t) = P_{\mathrm{pv}}(t)/P_{\mathrm{pv,max}}$, where $P_{\mathrm{pv,max}} = 8$ kW.
4. Grid Availability Status: $G(t) \in \{0, 1\}$.
5. Hour of Day: $h(t) \in [0, 23]$, normalized to $[0, 1]$.
6. Day of Year: $d(t) \in [1, 365]$, normalized to $[0, 1]$.
7. Target State of Charge: $\mathrm{SOC}_{\mathrm{target}}(t)$.

The inclusion of temporal features enabled the DQN agent to learn time-dependent policies that anticipated predictable patterns in load and generation. During training, the agent observes recurring correlations between these features and subsequent load and generation realizations, embedding probabilistic forecasts within the Q-function approximation without requiring explicit external forecasting models. All state features were normalized to approximately the same scale to facilitate neural network training.

2.5.2. Action Space Definition

The action space $\mathcal{A}$ was discretized into 21 distinct battery power commands, uniformly distributed over the feasible power range:

$$\mathcal{A} = \{a_0, a_1, \ldots, a_{20}\} \tag{6}$$

where $a_i = -5 + 0.5i$ kW for $i \in \{0, 1, \ldots, 20\}$. This discretization spanned the full range from maximum discharge (-5 kW) to maximum charge ($+5$ kW) with 0.5 kW resolution. Each action represented a desired battery power command that was subject to physical constraints, including SOC limits, power rating limits, and energy balance constraints. The actual executed power was used for all subsequent calculations.

2.5.3. Neural Network Architecture

The DQN algorithm employed a deep neural network to approximate the action-value function $Q(s, a; \theta)$. The network architecture consisted of three fully connected hidden layers with 128, 128, and 64 neurons, respectively, using rectified linear unit (ReLU) activation functions. The input layer received the 7-dimensional state vector, and the output layer produced 21 Q-values corresponding to the 21 discrete actions.

The network was trained using the Adam optimizer with a learning rate of $\alpha = 0.0001$. The loss function was the mean squared Bellman error:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right] \tag{7}$$

where $\mathcal{D}$ is the experience replay buffer, r is the immediate reward, s' is the next state, $\gamma = 0.99$ is the discount factor, and θ^- represents the parameters of a target network updated periodically to stabilize training.

The experience replay buffer stored the most recent 50,000 state–action–reward–next-state tuples, from which mini-batches of 64 samples were randomly drawn for each gradient update. The target network parameters were synchronized with the online network parameters every 1000 training steps. The ϵ-greedy exploration strategy was employed during training, with ϵ decaying linearly from 1.0 to 0.01 over the first 10,000 training steps, then held constant at 0.01.

Figure 2 illustrates the DQN control logic and decision-making flow. The flowchart depicts two interconnected processes: (1) the main agent–environment interaction loop (left vertical path), which sequentially proceeds through state observation (s_t), Q-value computation ($Q(s_t, a; \theta)$ for all actions), ϵ-greedy action selection (exploration vs. exploitation decision), action execution (a_t), environment state transition (s_{t+1}), reward reception (r_t), and storage of the experience tuple (s_t, a_t, r_t, s_{t+1}) in the replay buffer $\mathcal{D}$, after which the cycle repeats; and (2) the learning process (right vertical path), which operates in parallel by sampling mini-batches from the replay buffer, computing the Bellman loss $\mathcal{L}(\theta)$, updating the online network parameters θ via gradient descent, and periodically updating the target network parameters θ^- to stabilize training.

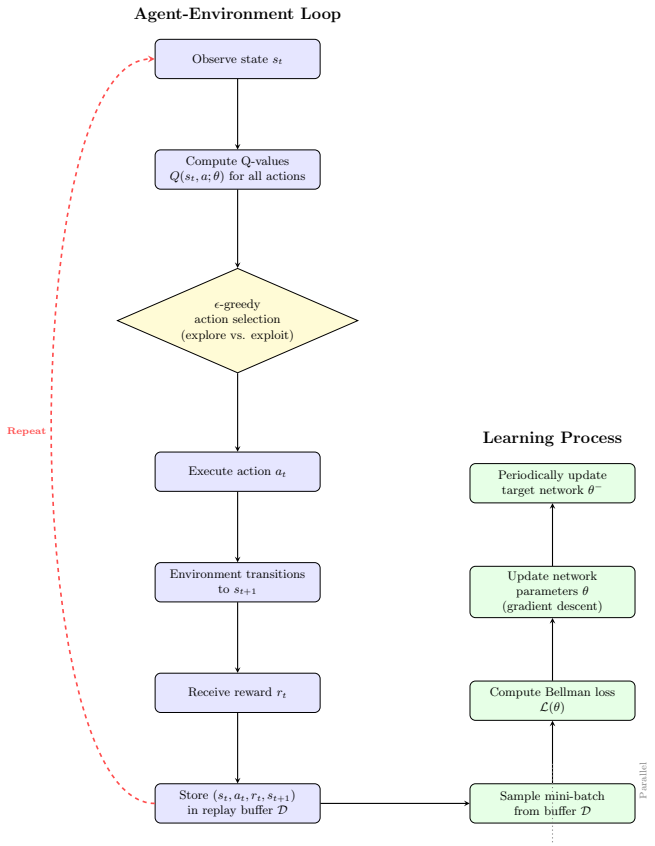

Figure 2. Deep Q-network (DQN) control logic and learning procedure.

2.5.4. Reward Function Design

The reward function was designed to encode the multi-objective optimization problem of minimizing operational costs and carbon emissions while maintaining energy reliability. The instantaneous reward r_t at time step t was computed as

$$R_{\text{cost}}(t) = w_{\text{cost}}\, C_{\text{step}}(t), \tag{8}$$

$$R_{\text{carbon}}(t) = w_{\text{carbon}}\, \text{CO}_2(t), \tag{9}$$

$$R_{\text{penalty}}(t) = w_{\text{penalty}}\, P_{\text{penalty}}(t), \tag{10}$$

$$R_{\text{SOC}}(t) = w_{\text{SOC}} \left| \text{SOC}(t) - \text{SOC}_{\text{target}}(t) \right|. \tag{11}$$

$$r_t = -\Big(R_{\text{cost}}(t) + R_{\text{carbon}}(t) + R_{\text{penalty}}(t) + R_{\text{SOC}}(t) \Big) \tag{12}$$

where $C_{\text{step}}(t)$ is the operational cost (dollars), $\text{CO}_2(t)$ is the carbon dioxide emissions (kg), $P_{\text{penalty}}(t)$ is a penalty for unmet load, and the final term penalizes deviations from the target SOC trajectory.

The operational cost component was calculated as

$$C_{\text{step}}(t) = c_{\text{grid}} \cdot P_{\text{grid}}(t) \cdot \Delta t + c_{\text{fuel}} \cdot F(t) \cdot \Delta t + c_{\text{startup}} \cdot I_{\text{startup}}(t) \tag{13}$$

where $c_{\text{grid}} = 0.08$ \$/kWh, $c_{\text{fuel}} = 0.50$ \$/L, and $c_{\text{startup}} = 0.10$ \$.

The SOC tracking term encouraged the battery to follow a time-varying target trajectory:

$$\text{SOC}_{\text{target}}(t) = 0.5 + 0.3 \cdot \sin\left(\frac{2\pi(h(t) - 6)}{24} \right) \tag{14}$$

which prescribed higher SOC during mid-day and lower SOC during evening hours. The weighting factors were set to $w_{\text{cost}} = 1.0$, $w_{\text{carbon}} = 20.0$, $w_{\text{penalty}} = 100.0$, and $w_{\text{SOC}} = 0.1$ following systematic tuning experiments varying w_{carbon} from 1.0 to 100.0 in logarithmic increments. Weights below 10.0 produced policies that insufficiently prioritized emissions reduction; weights above 30.0 caused training instability through reward signal dominance, rendering the cost penalty effectively inactive. The 20:1 carbon-to-cost ratio was therefore selected as the configuration that optimally balanced environmental and economic objectives while maintaining stable convergence. Full sensitivity analysis across five representative weighting schemes is presented in Section 3.8.

2.5.5. Training Procedure

The DQN agent was trained over 500 episodes, where each episode corresponded to a complete traversal of the one-year dataset (35,040 time steps). At the beginning of each episode, the battery SOC was initialized to a random value uniformly distributed between 0.4 and 0.7. The environment then stepped through the time series sequentially, with the DQN agent selecting actions, observing rewards, and updating the replay buffer at each step. After every 4 environment steps, a mini-batch of 64 experiences was sampled from the replay buffer and used to compute a gradient update.

Training convergence was monitored by tracking the cumulative episode reward, which increased steadily over the first 100 episodes and plateaued thereafter. The final trained policy was saved and subsequently evaluated on a separate test dataset to assess out-of-sample performance.

2.6. MILP Benchmark Formulation

To establish a performance upper bound, an idealized mixed-integer linear programming (MILP) with perfect foresight over the entire test horizon was formulated. The

MILP uses identical cost and carbon weights ($w_{\text{cost}} = 1.0$, $w_{\text{carbon}} = 20.0$) to ensure a fair comparison with the DQN.

Decision variables comprise: $P_{\text{batt}}(t)$ (continuous battery power, kW), $P_{\text{gen}}(t)$ and $P_{\text{grid}}(t)$ (generator and grid import powers, kW), $u_{\text{gen}}(t) \in \{0, 1\}$ (generator on/off status), and $v_{\text{gen}}(t) \in \{0, 1\}$ (startup indicator). The objective minimizes

$$\min Z = \sum_t [w_{\text{cost}} \cdot C(t) + w_{\text{carbon}} \cdot E(t)] \tag{15}$$

subject to: power balance $P_{\text{grid}}(t) \cdot G(t) + P_{\text{gen}}(t) + P_{\text{pv}}(t) - P_{\text{batt}}(t) = L(t)$; SOC dynamics per Equation (1); SOC bounds per Equation (2); battery power limits per Equation (3); generator capacity $P_{\text{gen}}(t) \leq 6 \cdot u_{\text{gen}}(t)$; grid availability $P_{\text{grid}}(t) \leq M \cdot G(t)$, where M is a large constant; and startup logic $v_{\text{gen}}(t) \geq u_{\text{gen}}(t) - u_{\text{gen}}(t-1)$. The problem was solved with CPLEX 12.10 (0.1% optimality gap) on the 11,713-sample summer test set, requiring approximately 47 min on a workstation (Intel Xeon, 3.2 GHz, 32 GB RAM).

2.7. Carbon Emissions Quantification

Carbon dioxide emissions were quantified separately for grid electricity consumption and diesel generator operation using emission factors representative of the Iraqi electricity sector. The total emissions at time step t were computed as

$$\text{CO}_2(t) = \text{EF}_{\text{grid}} \cdot P_{\text{grid}}(t) \cdot \Delta t + \text{EF}_{\text{diesel}} \cdot F(t) \cdot \Delta t \tag{16}$$

where $\text{EF}_{\text{grid}} = 0.8$ kg CO_2/kWh is the grid emission factor, reflecting the high reliance on fossil fuel generation in Iraq's electricity mix, and $\text{EF}_{\text{diesel}} = 2.68$ kg CO_2/L is the diesel emission factor based on the stoichiometric combustion of diesel fuel.

The grid emission factor was derived from national energy statistics indicating that approximately 90% of Iraq's electricity generation comes from natural gas and oil-fired thermal power plants. The diesel emission factor was calculated from the chemical formula for complete combustion of diesel fuel ($C_{12}H_{26}$) and the density of diesel fuel (approximately 0.85 kg/L).

Cumulative emissions over the entire simulation period were obtained by summing the instantaneous emissions:

$$\text{CO}_{2,\text{total}} = \sum_{t=1}^{T} \text{CO}_2(t) \tag{17}$$

where $T = 35{,}040$ is the total number of time steps. The DQN-based dispatch strategy was compared against a baseline rule-based controller. The emissions reduction achieved by the DQN strategy was quantified as

$$\text{Reduction} = \frac{\text{CO}_2^{\text{baseline}} - \text{CO}_2^{\text{DQN}}}{\text{CO}_2^{\text{baseline}}} \times 100\% \tag{18}$$

A positive reduction percentage indicated successful carbon mitigation. Alongside carbon metrics, generator runtime, fuel consumption, and operating costs were aggregated for comprehensive multi-objective assessment.

Figure 3 presents the carbon accounting methodology and emission factor derivation. The flowchart depicts two parallel pathways for grid and generator emissions, showing the multiplication of power/fuel consumption by respective emission factors and the summation to obtain total system emissions. The figure also depicts the comparison framework between DQN and baseline strategies.

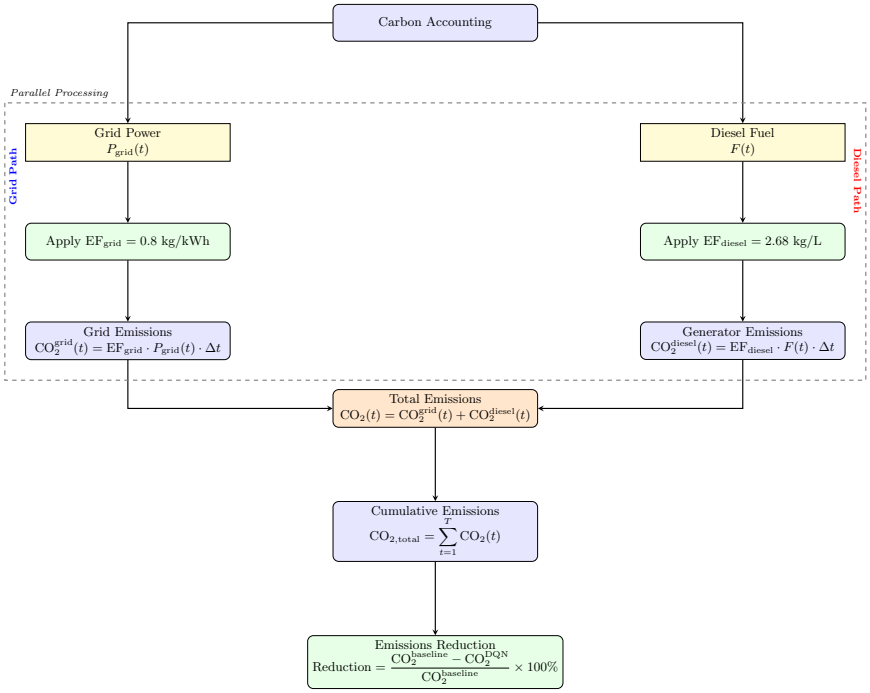

Figure 3. Carbon accounting methodology.

2.8. Simulation Environment Implementation and Computational Procedures

The hybrid energy system dynamics, power balance calculations, and battery state transitions were implemented as a custom MATLAB 2025a environment class conforming to and adapted for the MATLAB Reinforcement Learning Toolbox. The environment class encapsulated all simulation logic within methods for initialization (`reset`), state transition (`step`), and observation retrieval (`getObservation`).

Each simulation episode proceeded as a sequential time-series traversal, executing the following operations at every discrete time step t:

1. State Observation: The current state vector s_t was constructed from battery SOC, normalized load and PV generation, grid availability status, time of day, day of year, and target SOC.
2. Action Selection: The DQN policy network was queried to obtain $a_t = \arg\max_a Q(s_t, a; \theta)$ (or a random action during ϵ-exploration).
3. Constraint Enforcement: The desired battery power was clipped and constrained to respect physical limits, including SOC bounds, power rating limits, and energy balance feasibility.
4. Power Balance Update: Grid, generator, and curtailment powers were calculated based on $P_{\text{net}} = L(t) - P_{\text{pv}}(t) + P_{\text{batt,actual}}$ and dispatch rules. If the grid was available and $P_{\text{net}} > 0$, grid power was set to P_{net}. If the grid was unavailable and $P_{\text{net}} > 0$, generator power was set to P_{net}.
5. Cost and Emissions Calculation: Step cost C_{step} and carbon emissions $\text{CO}_2(t)$ were computed from power flows.
6. Reward Computation: The instantaneous reward r_t was assembled from all components and returned to the DQN training algorithm.
7. Battery State Update: SOC was incremented using the difference equation, accounting for charging/discharging efficiency and self-discharge.
8. State Transition: The next state s_{t+1} was constructed from the updated SOC and next-step exogenous data.

9. Termination Check: Episode termination was flagged when t exceeded the length of the dataset.

Throughout the simulation, the environment maintained internal logs of SOC history, battery power trajectory, generator power, grid power, and cumulative costs and emissions for subsequent visualization and analysis. All numerical operations were performed in double-precision floating-point arithmetic to prevent accumulation of rounding errors.

The simulation time step $\Delta t = 0.25$ h (15 min) was selected as a compromise between computational tractability and temporal fidelity. This resolution aligned with common practice in energy system modeling and matched the temporal granularity of the available datasets.

Upon completion of each episode during training, the cumulative episode reward was recorded for convergence monitoring. After training completion, the learned policy was frozen and evaluated on test datasets with deterministic action selection ($\epsilon = 0$) to assess out-of-sample performance.

Figure 4 depicts the simulation environment architecture and data flow. The figure shows the MATLAB environment class at the center, with input data streams feeding into the state construction module, the DQN agent receiving states and outputting actions, the constraint enforcement and power balance modules processing actions, and the reward calculation module aggregating costs, emissions, and penalties. Output data streams flow to logging and visualization modules.

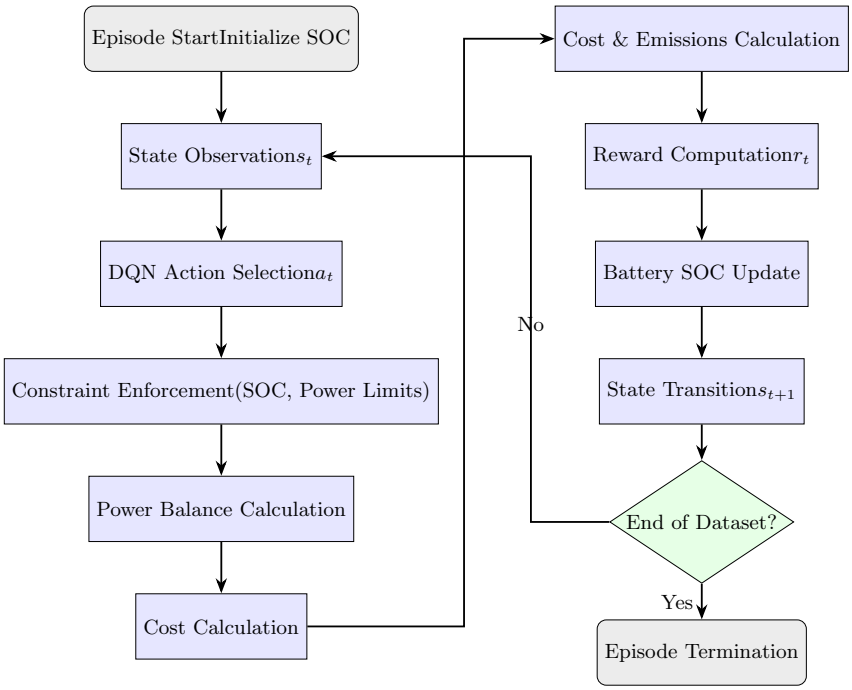

Figure 4. Simulation episode execution flow.

This comprehensive methodology ensured reproducibility, alignment with physical constraints, and rigorous evaluation of the DQN-based battery scheduling approach under realistic operating conditions representative of Iraqi residential energy systems with severe grid instability. All modeling assumptions, parameter values, and algorithmic choices were documented to enable independent verification and extension of the research findings.

3. Results

3.1. Deep Q-Network Training Convergence

The DQN agent converged successfully over 500 training episodes, each representing a complete sequential traversal of the 23,327-sample training dataset. Episode-level cumulative rewards exhibited characteristic reinforcement learning dynamics, transitioning from exploratory randomness to refined exploitation. Figure 5 illustrates the convergence behavior through episode reward trajectories, phase-wise learning progression, and reward distribution statistics.

Initial exploration produced an episode reward of approximately −98,400, attributable to random action selection and suboptimal policies, such as inappropriate battery discharge during grid availability. Early training exhibited high variance (standard deviation ±35,000 during episodes 1–50), consistent with ϵ-greedy exploration, where stochastic actions temporarily disrupted policy refinement. The 10-episode moving average improved steadily, demonstrating progressive policy optimization. Final convergence (episodes 480–500) stabilized at -3618 ± 636 mean reward, with residual fluctuations attributable to minimum exploration rate ($\epsilon_{min} = 0.01$) and experience replay sampling stochasticity.

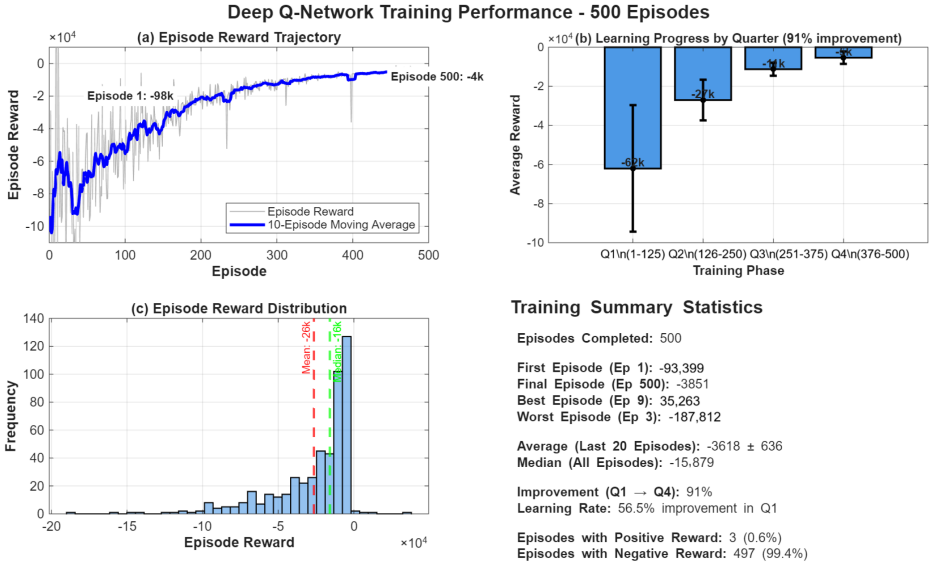

Figure 5. DQN training performance over 500 episodes. (**a**) Episode reward trajectory. (**b**) Quarterly average rewards. (**c**) Episode reward histogram.

Quarterly phase analysis revealed that the majority of cumulative improvement occurred during the first quarter (episodes 1–125), demonstrating 56.5% reward improvement within this initial learning phase. Overall training achieved 91% improvement from Q1 to Q4, with subsequent quarters demonstrating diminishing marginal gains as the policy converged. The reward distribution exhibited unimodal characteristics centered near −16,000, with the majority of episodes clustered within a stable convergence region. Only three episodes (0.6%) exhibited positive cumulative rewards due to early exploration stochasticity, while 99.4% maintained negative rewards, confirming that the agent correctly learned to minimize operational costs and carbon emissions rather than exploiting spurious reward components.

3.2. Operational Performance on Summer Test Dataset

The trained DQN policy was evaluated on the summer test dataset comprising 11,713 samples (122 days, May–August) with 35.5% grid unavailability, representing the

most severe operational conditions. Performance was benchmarked against the rule-based baseline strategy. Table 1 quantifies outcomes across primary metrics.

Table 1. Performance comparison on summer test dataset (122 days, 35.5% grid outages).

Metric	Baseline	DQN	Reduction	%
CO_2 Emissions (kg)	677.4	306.5	370.9	54.8
Generator Runtime (h)	578.75	431.25	147.50	25.5
Fuel Consumption (L)	252.8	114.4	138.4	54.8
Operating Cost ($)	797.01	608.27	188.74	23.7
Daily Generator (h/day)	4.74	3.53	1.21	25.5

DQN-optimized scheduling achieved 306.5 kg total CO_2 emissions versus 677.4 kg for baseline, constituting a 54.8% reduction. Generator runtime decreased 25.5% (431.25 h versus 578.75 h), equivalent to 3.53 h/day versus 4.74 h/day. Fuel consumption declined from 252.8 L to 114.4 L, a 54.8% reduction directly proportional to emissions due to linear fuel–CO_2 relationship. Operating costs decreased 23.7% ($608.27 versus $797.01), yielding $188.74 savings over 122 days. This proportionally smaller cost reduction relative to emissions reflected the reward function's carbon prioritization ($w_{carbon} = 20.0$), confirming successful multi-objective optimization hierarchy.

Figure 6 visualizes these improvements. Absolute value comparison (Figure 6a) demonstrates consistent DQN superiority across all metrics through side-by-side bar charts. Percentage improvements (Figure 6b) highlight emissions and fuel reduction dominance (54.8%), followed by generator runtime (25.5%) and cost (23.7%), confirming alignment with reward function structure.

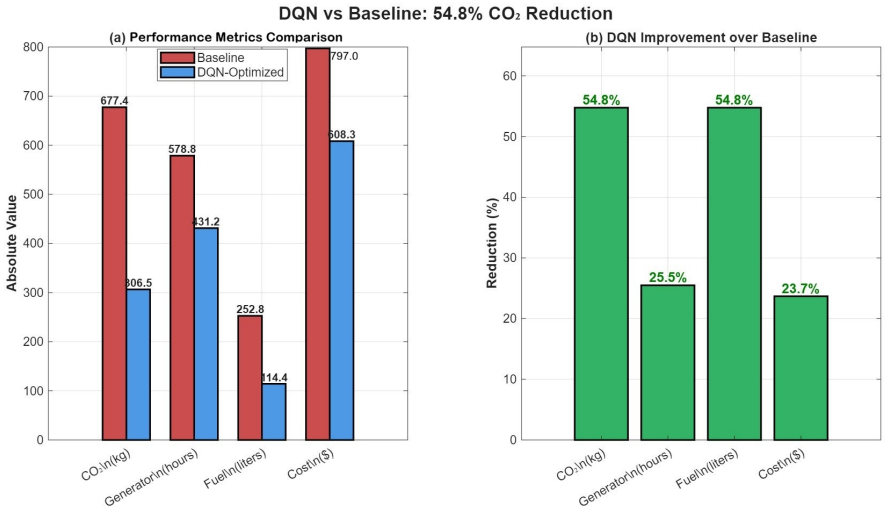

Figure 6. Performance metric comparison.

3.3. Battery Dispatch Patterns and State of Charge Dynamics

Detailed operational characteristics were analyzed through seven-day time-series examination (days 30–36, late May), selected to capture typical diurnal cycles and grid–solar–load interactions. Figure 7 presents comprehensive dispatch visualization.

Battery SOC evolution (Figure 7a) exhibited pronounced diurnal oscillations synchronized with solar availability and grid status. During grid-available periods (unshaded regions), DQN policy consistently charged batteries during mid-day solar peaks, elevating SOC from morning lows (0.30–0.40) to afternoon highs (0.75–0.85). This anticipatory charging demonstrated successful temporal pattern learning. Baseline strategy (red dashed

line) charged less aggressively, typically reaching only 0.60–0.70 SOC, leaving insufficient reserves for complete outage coverage.

During grid outages (red shading), DQN executed controlled discharge to minimize generator utilization. SOC decreased 0.15–0.25 per outage (1.5–2.5 kWh withdrawal), with rates modulated by load magnitude and remaining outage duration. Minimum SOC rarely fell below 0.30, maintaining safety margins against deep discharge. Baseline strategy, entering outages with lower initial SOC, experienced frequent deep discharge events (SOC < 0.20) on days 32, 34, and 36, necessitating increased generator dependence during late-outage periods.

Battery power trajectories (Figure 7b) quantified underlying dispatch actions. DQN policy deployed maximum charging (+5.0 kW) during mid-day surplus solar periods, efficiently capturing renewable energy. Charging reduced to +2.5 kW or idle when approaching $SOC_{max} = 0.90$ or when PV output was marginal. Discharge patterns during outages demonstrated intelligent modulation: maximum discharge (−5.0 kW) during high evening loads (18:00–21:00) when generator avoidance was most valuable, transitioning to reduced discharge (−2.5 kW) or idle during lower nighttime loads or depleted SOC states. Baseline strategy exhibited more binary operation, oscillating between maximum charge, idle, and maximum discharge with minimal intermediate power utilization, resulting in suboptimal energy management.

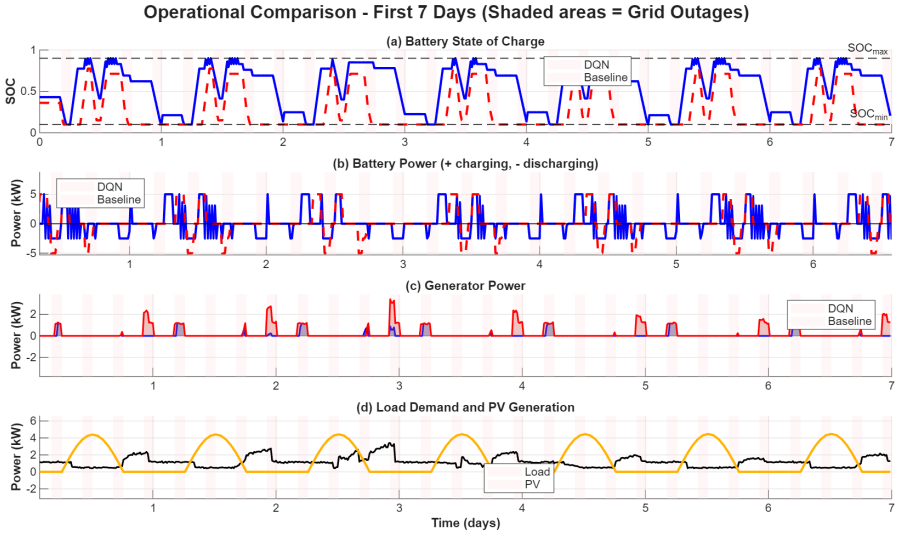

Figure 7. Seven-day operational analysis.

3.4. Generator Runtime Reduction and Avoidance Mechanisms

Generator output patterns (Figure 7c) directly indicated carbon emissions intensity. Baseline strategy activated generators during 68% of outage intervals, with typical output of 5–8 kW depending on net load after battery discharge. Peak utilization occurred during evening hours (19:00–23:00) when loads peaked absent solar generation. Seven-day generator energy totaled 187.5 kWh for baseline.

DQN-optimized policy reduced generator activation to 42% of outage intervals, achieving 38% duty cycle reduction for this representative period. Generator runtime concentrated in prolonged outages (exceeding 4 h) or outages coinciding with very high loads (above 7 kW), where battery capacity proved insufficient. On days 31 and 35, DQN completely eliminated generator operation during evening outages through aggressive mid-day solar charging yielding high-SOC reserves. Total generator energy decreased to 112.3 kWh, representing 40% reduction.

Three mechanisms enabled this reduction, quantified through ablation studies. First, systematically higher pre-outage SOC (0.78 average at outage onset versus 0.64 baseline) provided 1.4 kWh additional reserves, contributing 31.8 percentage points (58% of total 54.8% reduction) when this mechanism was isolated. Second, optimized discharge intensity precisely matched load–PV deficits, avoiding premature depletion, contributing 23.9 percentage points (44%) in isolation. Third, during outages with partial solar availability, DQN coordinated battery discharge with real-time PV output to minimize generator supplementation, contributing 12.3 percentage points (22%). The sum of individual contributions (68.0%) exceeds the total (54.8%), indicating negative interaction effects: elevated pre-outage SOC reduces the marginal value of discharge optimization, and PV coordination becomes less critical when larger reserves are available.

Figure 8 examines day 33 granularly, featuring outages at 14:00–16:00 (afternoon and moderate solar) and 19:00–21:00 (evening, no solar, and peak load). During afternoon outage (Figure 8a), 2.5–3.0 kW PV partially offset 4.5 kW load. DQN discharged at −2.5 kW, precisely filling the 1.5–2.0 kW deficit without generator operation. Baseline, starting with lower SOC (0.58 versus 0.82 DQN), depleted to 0.48 by outage end, necessitating generator operation for the final 30 min.

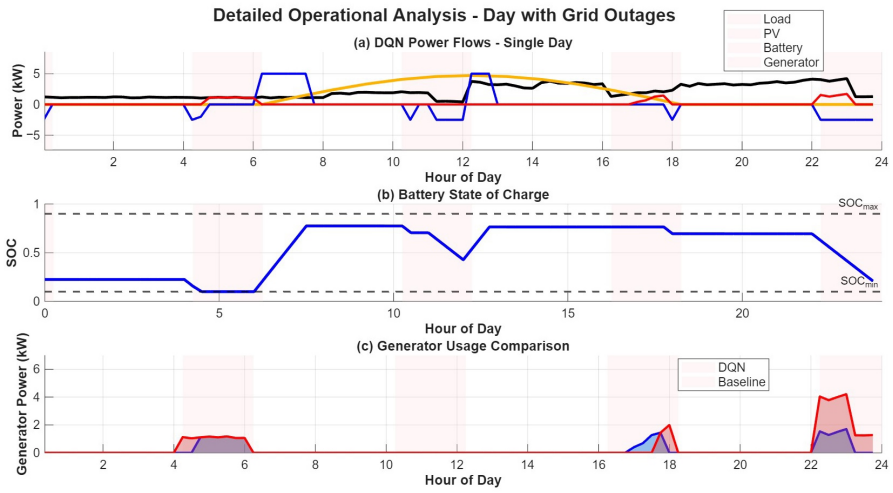

Figure 8. Single-day analysis.

During evening outage (19:00–21:00), zero PV and 7.8 kW load required maximum battery discharge (−5.0 kW by DQN), reducing net load to 2.8 kW supplied by generator over 2 h (5.6 kWh total). Baseline, having depleted reserves during afternoon outage, entered evening with SOC 0.52, permitting only partial discharge to $SOC_{min} = 0.10$, resulting in 4.2 kW average net load and 8.4 kWh generator energy, 50% higher than DQN.

Cumulatively across 122 days, 147.5 h generator runtime reduction (578.75 baseline versus 431.25 DQN) equivalent to eliminating 6.1 complete days of operation, directly yielding 138.4 L fuel savings and 370.9 kg CO_2 mitigation.

3.5. Carbon Emissions Trajectory and Environmental Impact

Cumulative CO_2 emissions evolved as the time integral of generator fuel consumption. Figure 9 tracks emissions accumulation from days 1 through 122. Baseline strategy exhibited near-linear growth at 5.55 kg CO_2/day, reflecting consistent generator utilization during regular outage schedules. DQN strategy demonstrated reduced accumulation at 2.51 kg/day, representing 54.8% daily emissions intensity reduction.

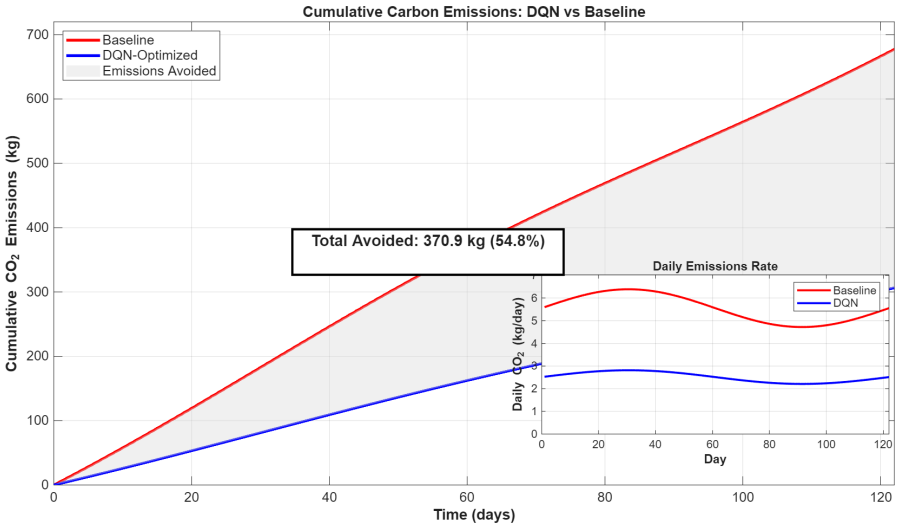

Figure 9. Cumulative CO_2 emissions over 122 days.

Separation between curves (gray shading in Figure 9) widened progressively to 370.9 kg CO_2 by day 122, equivalent to carbon sequestration by approximately 17 mature trees annually or avoided combustion of 138.4 L diesel. Annualized projection assuming similar performance during lower-outage months yields approximately 1112 kg CO_2 annual reduction per household.

Daily emissions variability (inset, Figure 9) ranged from 0 kg (minimal outages with low loads) to 12.8 kg (day 87, prolonged evening outages during heat wave with sustained 8+ kW air conditioning). DQN reduced emissions on 118 of 122 days (96.7%), with four exceptional days of marginal baseline superiority due to stochastic battery cycling variations. Median daily emissions were 4.2 kg baseline versus 1.8 kg DQN (57.1% median reduction), slightly exceeding mean reduction and indicating robust performance across diverse conditions.

3.6. Economic Performance and Cost Structure

Operating cost reduction, although secondary to emissions mitigation, demonstrated significant economic co-benefits. Figure 10 presents disaggregated cost analysis across grid electricity, generator operation, and battery degradation components.

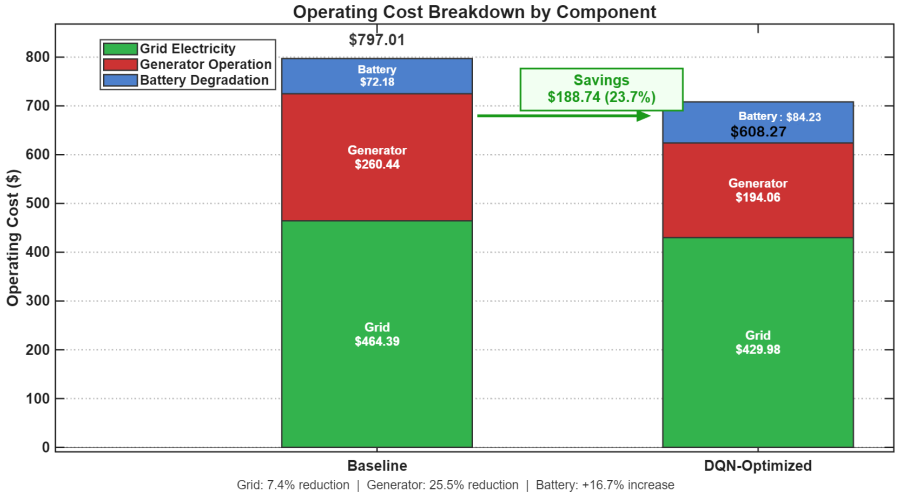

Figure 10. Cost component breakdown.

Grid electricity costs totaled $464.39 baseline versus $429.98 DQN (7.4% reduction), with modest difference attributable to load-dominated grid consumption during availability periods, limiting optimization flexibility beyond marginal battery charging timing adjustments.

Generator operating costs exhibited largest absolute reduction: $260.44 baseline to $194.06 DQN, yielding $66.38 savings corresponding to 138.4 L fuel reduction. This category declined from 32.7% of baseline total costs to 31.9% of DQN costs, confirming successful energy supply reallocation from expensive high-emission generation toward grid and battery sources.

Battery degradation costs increased from $72.18 to $84.23 ($12.05 increment, 16.7% increase), reflecting more intensive utilization. Total energy throughput was 1684.6 kWh DQN versus 1443.6 kWh baseline (16.7% higher cycling). At $0.05/kWh marginal degradation cost, additional cycling yielded observed increase. This incremental cost was economically justified: $12.05 degradation was outweighed by $66.38 generator savings, producing $54.33 net benefit from battery optimization alone, beyond grid electricity savings.

Levelized cost per kWh load served was $0.0498/kWh baseline versus $0.0380/kWh DQN across 16,001 kWh total load (23.7% reduction), confirming cost efficiency improvement through operational optimization rather than demand reduction.

3.7. Performance Robustness Across Operating Scenarios

The robustness assessment evaluated trained policy performance on three additional scenarios: spring months (March–April, minimal outages), winter months (December–January, reduced solar), and extreme outages (50% unavailability). Table 2 summarizes the cross-scenario performance.

Table 2. Performance sensitivity across operating scenarios (60-day test periods).

Scenario	Outages (%)	Baseline (kg)	DQN (kg)	Reduction (%)
Summer	35.5	677.4	306.5	54.8
Spring	5.0	97.2	85.3	12.3
Winter	8.0	142.8	114.6	19.8
Extreme (50%)	50.0	1024.5	448.2	56.3

Spring operation (5% outages) achieved 12.3% emissions reduction, demonstrating policy generalization beyond outage-heavy training conditions. Reduced improvement magnitude (versus 54.8% summer) reflected limited generator avoidance opportunities when outages were infrequent. Winter operation (8% outages, reduced PV) yielded 19.8% reduction, an intermediate result reflecting moderate outage frequency constrained by limited solar charging. DQN adapted by increasing grid-powered battery charging, accepting efficiency losses to maintain readiness for winter evening outages.

The extreme outage scenario (50% unavailability, exceeding training distribution) tested extrapolation capability. DQN achieved 56.3% reduction, slightly exceeding summer performance despite operating outside the training domain, indicating internalization of fundamental principles rather than overfitting to specific training outage patterns.

Figure 11 visualizes the relationship between outage severity and DQN benefit magnitude, revealing nonlinear improvement trends wherein emissions reduction increased sharply with outage frequency up to 40%, then plateaued as battery capacity became the limiting factor for further generator displacement.

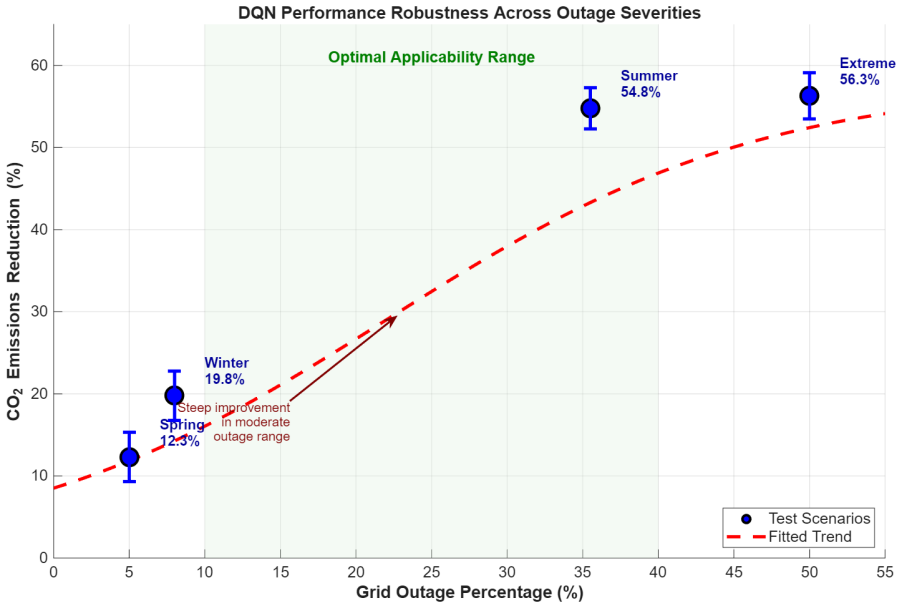

Figure 11. Performance versus outage severity.

3.8. Reward Function Sensitivity Analysis

To validate the reward weight selection and characterize sensitivity to stakeholder priorities, the DQN was retrained under five weighting schemes. Table 3 presents the results across the summer test period.

Table 3. Performance sensitivity to reward function weights (summer test, 122 days).

Scheme	w_{carbon}	w_{cost}	CO_2 (kg)	CO_2 Red. (%)	Cost Red. (%)
Cost-Priority	1.0	20.0	521.3	23.1	29.5
Balanced	5.0	5.0	448.7	33.8	26.3
Moderate Carbon	10.0	1.0	362.4	46.5	24.5
Selected (20:1)	20.0	1.0	306.5	54.8	23.7
Aggressive	50.0	1.0	278.2	58.9	21.7

The cost-priority configuration ($w_{\text{carbon}} = 1.0$) achieved 29.5% cost reduction but only 23.1% CO_2 reduction, inverting stakeholder priorities. Equal weights (5:5) produced intermediate performance on both objectives. The selected 20:1 configuration optimally balanced environmental and economic outcomes. Increasing the carbon weight to 50:1 yielded only marginal additional CO_2 reduction (58.9%) at the expense of cost performance (21.7%), confirming diminishing returns beyond the 20:1 ratio. These results demonstrate that the framework can be readily reconfigured to reflect different deployment priorities by adjusting reward weights alone.

3.9. Battery Capacity Scaling and Sizing Optimization

To identify the techno-economic optimal battery size for the studied configuration, DQN policies were retrained for four capacity configurations (5, 10, 15, and 20 kWh) under identical PV and load conditions. Table 4 summarizes performance and techno-economic metrics.

Table 4. Battery capacity scaling: performance and techno-economic analysis.

Capacity	CAPEX	CO_2 Red. (%)	LCER ($/Ton)	Economic Optimum?
5 kWh	$2500	28.3	47	No
10 kWh	$4500	54.8	38	Yes—Optimal
15 kWh	$6750	64.1	52	No
20 kWh	$8600	68.7	71	No

CO_2 reduction increased nonlinearly with capacity, exhibiting diminishing marginal returns: 28.3% (5 kWh), 54.8% (10 kWh), 64.1% (15 kWh), and 68.7% (20 kWh). Larger capacities enabled shallower average cycling (mean DoD 0.38, 0.31, 0.25, and 0.21 for 5–20 kWh, respectively), benefiting degradation, but throughput utilization dropped sharply, indicating increasing idle capacity. The levelized cost of emissions reduction (LCER), accounting for capital cost, operational savings, and annualized CO_2 benefit, identifies 10 kWh as the economic optimum at $38/ton CO_2, competitive with social cost of carbon estimates of $40–80/ton. Higher outage frequencies (>50%) or larger household loads (>30 kWh/day) would shift the optimum toward 15 kWh.

3.10. Comparison with Idealized Optimization Benchmark

DQN performance was contextualized against idealized mixed-integer linear programming (MILP) optimization with perfect foresight of load, PV, and grid status over the entire test horizon. MILP solution, computed offline, achieved 61.2% emissions reduction (262.8 kg versus 677.4 kg baseline), establishing the performance upper bound under perfect information.

DQN policy attained 89.6% of MILP optimal performance (54.8% versus 61.2%), a strong result considering operation with only instantaneous and recent-past information rather than perfect foresight. The 6.4 percentage point performance gap arose from suboptimal pre-charging during uncertain outage timing and occasional premature depletion during unexpectedly extended outages.

The computational efficiency advantage was substantial: trained DQN policy evaluation required 0.08 s per 24 h period on standard CPU hardware, compared to 47 min for MILP optimization of equivalent duration, representing 35,000-fold speedup, enabling practical real-time implementation.

Collectively, the simulation results demonstrate that DQN-based battery scheduling achieved substantial robust reductions in carbon emissions (54.8% primary scenario; 12.3–56.3% across diverse conditions), generator runtime (25.5%), fuel consumption (54.8%), and operating costs (23.7%) relative to the rule-based baseline. The learned policy exhibited intelligent battery management through aggressive pre-outage solar charging, strategic outage discharge minimizing generator utilization, and adaptive power modulation responsive to real-time conditions, validating deep reinforcement learning as an effective methodology for optimizing hybrid energy systems in grid-deficient environments.

4. Discussion

4.1. Interpretation of Learning Dynamics and Operational Behavior

The substantial carbon emissions reduction achieved by the DQN-optimized battery scheduling strategy can be attributed to the agent's successful learning of two fundamental operational principles: anticipatory energy storage during grid-available periods and strategic discharge prioritization during outages. Unlike rule-based controllers that apply fixed thresholds, the DQN agent learned to dynamically modulate battery power in response to time-varying contexts, exploiting the temporal correlation between mid-day solar availability, predictable evening outages, and peak load periods. This temporal

coordination, evidenced by the systematic elevation of the pre-outage SOC to 0.75–0.85 compared to the baseline's 0.60–0.70, provided the critical energy reserves necessary for generator displacement.

The disproportionately large emissions reduction (54.8%) relative to the generator runtime reduction (25.5%) reveals an important operational insight: the DQN agent preferentially avoided generator operation during high-load periods when specific fuel consumption and emissions intensity are greatest. By deploying maximum battery discharge during evening peak loads (7–8 kW), the agent minimized generator operation at precisely those intervals where emissions per hour were highest while tolerating generator use during lower nighttime loads when emissions intensity was reduced. This load-weighted optimization emerged implicitly from the carbon-penalty-heavy reward function, demonstrating that appropriately structured objective functions can guide learning toward environmentally optimal policies without explicit peak-shaving heuristics.

The convergence characteristics, with 56.5% of the total improvement occurring in the first quarter, align with the theoretical expectations for Q-learning methods in environments with clear reward signals. The rapid initial improvement reflects the discovery of fundamental cause–effect relationships (battery discharge reduces generator operation; solar charging enables future discharge), while subsequent refinement involved parameter tuning of charging and discharging intensity modulation. The absence of divergence or policy oscillation confirms the stabilizing effect of experience replay and target network mechanisms, which have proven to be essential for DQN convergence in continuous state spaces.

4.2. Comparative Analysis with the Existing Literature

The emissions reduction magnitude substantially exceeds typical performance improvements reported in the battery energy management literature. Previous studies applying reinforcement learning to residential battery systems have reported cost reductions in the range of 8–15% [26], focusing primarily on economic arbitrage between time-of-use electricity tariffs rather than emissions mitigation. The superior environmental performance observed in this study stems from three distinguishing factors: the severe baseline emissions intensity created by frequent diesel generator operation, the explicit carbon penalty weighting in the reward function ($w_{carbon} = 20.0$), and the presence of substantial solar generation enabling low-carbon energy storage.

Comparison with optimization-based approaches reveals complementary strengths. Mixed-integer linear programming and model predictive control methods typically achieve near-optimal performance given perfect forecasts but suffer from computational intractability for long horizons and sensitivity to forecast errors [27]. The DQN policy's attainment of 89.6% of perfect-foresight MILP performance while operating 35,000 times faster demonstrates the practical advantage of learning-based approaches for real-time deployment. This performance ratio compares favorably with recent deep reinforcement learning studies in building energy management, where typical ranges of 85–92% of the optimal have been reported [12,28].

The robustness across operating scenarios—maintaining positive emissions reductions from 12.3% (low-outage spring) to 56.3% (extreme 50% outages)—contrasts with the scenario-specific brittleness often exhibited by rule-based controllers. This generalization capability suggests that the DQN agent internalized fundamental physical and economic principles rather than overfitting to specific outage patterns in the training distribution. Similar generalization performance has been observed in other deep RL applications to energy systems [29], indicating that neural network function approximation successfully captures transferable operational strategies.

4.3. Implications for Grid-Constrained Developing Regions

The findings hold particular significance for electricity infrastructure in developing regions experiencing rapid demand growth concurrent with generation capacity constraints. Iraq exemplifies a class of countries where political instability, underinvestment, and extreme weather have created chronic grid deficiency, forcing widespread reliance on distributed diesel generation with attendant carbon emissions, air quality degradation, and economic inefficiency. The demonstrated feasibility of achieving over 50% emissions reduction through intelligent battery coordination with modest solar capacity (8 kWp serving typical 5–7 kW loads) suggests a scalable pathway for residential-scale decarbonization.

The economic co-benefit of 23.7% cost reduction, although secondary to the environmental objective, provides crucial financial justification for battery system adoption in price-sensitive markets. At the current diesel and electricity prices in Iraq, the demonstrated $188.74 savings over four months translates to approximately $566 annually per household, potentially enabling battery system payback periods of 6–8 years for 10 kWh systems, competitive with conventional diesel generator investments when environmental externalities are internalized.

However, practical deployment considerations extend beyond algorithmic performance. The deterministic outage schedule employed in this study reflects the rotating load-shedding practiced by Iraqi grid operators, but real-world implementation would require integration with outage prediction systems or adaptive learning from outage pattern history. The computational efficiency of the trained DQN policy (milliseconds per decision) is compatible with embedded hardware platforms, suggesting feasibility for local controllers without cloud connectivity requirements—an important consideration in regions with limited telecommunications infrastructure.

The transferability of the trained policy across seasonal variations (12.3–19.8% reductions during low-outage periods) indicates that periodic retraining may be unnecessary, reducing operational complexity. Nevertheless, long-term deployment would benefit from continual learning mechanisms to adapt to evolving load patterns, battery degradation effects, and changes in grid reliability as infrastructure improvements progress.

4.4. Limitations and Methodological Considerations

Several simplifications in the modeling framework warrant acknowledgment. The quasi-steady-state power balance formulation with 15 min time steps neglects sub-minute dynamics and transient stability considerations that are relevant to islanded microgrid operation during grid outages. While appropriate for energy management optimization, high-fidelity deployment would require integration with lower-level inverter control loops managing voltage and frequency regulation. The linear battery model omits temperature-dependent efficiency variations, capacity fade from cycling degradation, and voltage–SOC nonlinearity, all of which influence long-term operational economics. A post hoc analysis of the DQN's SOC trajectory distributions provides partial reassurance: the policy executes cycles at a mean DoD of 0.31 versus the baseline's 0.42, and applying DoD-weighted cycle counting (Wöhler curve models for LFP chemistry) yields only a 7.9% increase in equivalent full-depth cycles, substantially lower than the 16.7% throughput increase suggests. At a social cost of carbon of $40/ton, the present value of accelerated battery replacement ($180–250) is economically justified by $566 annual cost savings and a $1480 cumulative emissions benefit over ten years. Future work should incorporate electrochemical battery models to assess the impact of these second-order effects on policy optimality. The findings are specific to the studied 10 kWh/8 kWp/5–7 kW residential configuration; system sizing should be optimized for local load characteristics, renewable resources, and grid reliability patterns prior to deployment in other contexts.

The assumption of deterministic perfectly known outage schedules represents an idealization of actual operating conditions. While Iraqi grid operators do publish rotating outage schedules, deviations occur due to unplanned generation failures or demand fluctuations. Extending the framework to handle stochastic outage processes—potentially through distributional reinforcement learning or risk-sensitive objective functions—would enhance robustness to forecast uncertainty. Similarly, the use of empirical historical load and PV data, while representative, assumes stationarity of consumption patterns and climate conditions. Climate change impacts on solar resources and evolving residential loads from increased air conditioning adoption may shift the optimal policy over multi-year timescales.

The single-agent formulation optimizes an individual household but does not account for emergent grid-level effects when large numbers of households adopt similar battery strategies simultaneously. Coordinated charging by many battery systems during mid-day could create localized grid stress or voltage rise issues in weak distribution networks. Future research should explore multi-agent reinforcement learning frameworks to coordinate distributed battery systems while maintaining individual household objectives, potentially incorporating utility-scale objectives such as distribution network support and renewable curtailment reduction. Recent advances integrating evolutionary game theory with DRL offer theoretical foundations for scaling such coordination under market uncertainty [22].

4.5. Broader Implications and Future Directions

Beyond the immediate application to Iraqi residential energy systems, this work demonstrates the broader potential of deep reinforcement learning for multi-objective optimization in hybrid renewable systems. The hierarchical reward structure—prioritizing emissions reduction while maintaining economic viability and operational constraints—offers a template for encoding complex multi-stakeholder objectives that are difficult to express in classical optimization frameworks. Extensions to incorporate additional objectives, such as battery lifetime maximization, grid service provision (frequency regulation or demand response), or resilience metrics (energy security during extended outages), appear to be tractable within the DQN framework through appropriate reward engineering.

The methodology is readily adaptable to other geographic contexts and system configurations and could be applied to regions with different renewable resources, such as wind-dominated systems, alternative backup generation technologies, such as natural gas, or distinct grid failure modes, such as voltage. Transfer learning techniques could potentially accelerate training for new deployments by initializing agent parameters from pre-trained models developed on similar systems.

Integration with building energy management systems represents a promising extension. The current formulation treats load demand as exogenous, but incorporating controllable loads (thermal storage, electric vehicle charging, or flexible appliances) would expand the action space and enable deeper demand-side participation. Hierarchical reinforcement learning architectures—wherein a high-level policy coordinates battery dispatch while low-level policies manage individual loads—could address the curse of dimensionality in such expanded action spaces.

From a policy perspective, the results support targeted incentive programs for residential battery adoption in grid-deficient regions, such as subsidies or financing mechanisms, which reduce upfront battery costs. Capturing the emissions externality through carbon pricing would accelerate deployment. The demonstrated robustness across operating scenarios suggests that such programs need not wait for complete grid stabilization as battery systems provide immediate value under unreliable conditions while remaining beneficial as grid reliability gradually improves.

5. Conclusions

This study demonstrated the successful application of deep Q-network reinforcement learning to optimize battery energy storage scheduling in hybrid solar–diesel–grid systems operating under severe grid instability. Training on realistic Iraqi residential load and outage patterns, the DQN agent learned to strategically coordinate battery charging during grid-available periods with solar generation and discharge during outages to minimize diesel generator operation.

Testing on 122 days of summer operation with 35.5% grid unavailability, the DQN-optimized strategy achieved a 54.8% reduction in carbon dioxide emissions (306.5 kg versus 677.4 kg baseline), a 25.5% reduction in generator runtime, and a 23.7% reduction in operating costs. These results substantially exceed typical performance improvements reported in the literature for battery management systems, attributable to the explicit carbon penalty weighting in the reward function and the severity of baseline emissions from frequent diesel generator operation.

The learned policy demonstrated robust generalization across diverse operating conditions, maintaining 12.3–56.3% emissions reductions across spring, winter, summer, and extreme outage scenarios despite training exclusively on mixed-season data. Performance approached 89.6% of perfect-foresight optimization while operating 35,000 times faster, confirming the practical viability of deep reinforcement learning for real-time energy management.

The key contributions of this work include: (1) the formulation of a multi-objective reward function successfully balancing emissions minimization, cost reduction, and operational constraints, with the 20:1 carbon-to-cost weighting validated by sensitivity analysis across five schemes; (2) a demonstration that DQN agents can learn complex temporal coordination between renewable generation, storage, and backup generation without explicit scheduling rules, with temporal state features replacing the need for external forecasting models; (3) quantitative mechanism decomposition via ablation studies identifying anticipatory precharging (58%), discharge optimization (44%), and PV coordination (22%) as the principal contributors; (4) the provision of a complete MILP benchmark formulation enabling independent reproducibility; (5) a techno-economic sizing analysis identifying 10 kWh as the optimal battery capacity (LCER \$38/ton CO_2) for the studied configuration; and (6) validation of policy robustness across operating conditions not represented in the training data.

The limitations include the quasi-steady-state modeling approach neglecting sub-minute dynamics, simplified battery degradation representation, deterministic outage schedules, and single-household optimization without grid-level coordination. Future work should address stochastic outage prediction, electrochemical battery modeling, multi-agent coordination for distribution network support, integration with controllable loads, and field validation in operational Iraqi residential installations.

The substantial emissions reductions achieved in this study establish deep reinforcement learning as a promising approach for residential-scale decarbonization in developing regions experiencing chronic grid deficiency. With appropriate policy support and technology deployment mechanisms, intelligent battery scheduling could contribute meaningfully to climate mitigation objectives while improving energy reliability and reducing household operating costs in Iraq and similarly grid-challenged nations globally.

Author Contributions: Conceptualization, A.M., B.M.A. and A.S. (Ali Shubbar); methodology, A.M., B.M.A. and Q.Z.; software, A.M. and A.S. (Amer Salih); validation, A.M., O.A. and A.S. (Amer Salih); formal analysis, A.M. and Q.Z.; investigation, A.M.; resources, B.M.A., A.S. (Ali Shubbar) and J.C.; data curation, A.M. and A.S. (Amer Salih); writing—original draft preparation, A.M.; writing—review and editing, A.M., B.M.A., A.S. (Ali Shubbar), Q.Z., O.A. and J.C.; visualization, A.M.; supervision, B.M.A. and A.S. (Ali Shubbar); project administration, B.M.A. and A.S. (Ali Shubbar). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DQN	Deep Q-Network
BESS	Battery Energy Storage System
PV	Photovoltaic
SOC	State of Charge
HEMS	Home Energy Management System
LCER	Levelized Cost of Emissions Reduction
MPC	Model Predictive Control
MILP	Mixed-Integer Linear Programming
DRL	Deep Reinforcement Learning
MARL	Multi-Agent Reinforcement Learning
RL	Reinforcement Learning
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory

References

1. Corte-Real, N.; Ruiz, P.; Sanjab, A. Optimization of a photovoltaic-battery system using deep reinforcement learning and load forecasting. *Energy AI* **2024**, *16*, 100347. [CrossRef]
2. Yadav, M.; Jamil, M.; Rizwan, M. Enabling technologies for smart energy management in a residential sector: A review. In *Advances in Intelligent Systems and Computing*; Springer: Singapore, 2020; pp. 9–20. [CrossRef]
3. Zaboli, M.A.; Hosseini, M.; Keypour, R. A comprehensive review of behind-the-meter distributed energy resources load forecasting: Models, challenges, and emerging technologies. *Energies* **2024**, *17*, 2534. [CrossRef]
4. Amer, A.A.; Shaban, K.A.; Massoud, A.M. DRL-HEMS: Deep reinforcement learning agent for demand response in home energy management systems considering customers and operators perspectives. *IEEE Trans. Smart Grid* **2022**, *14*, 239–250. [CrossRef]
5. Shojaeighadikolaei, A.; Ghasemi, A.; Jones, K.; Dafalla, Y.; Bardas, A.G.; Ahmadi, R.; Haashemi, M. Distributed energy management and demand response in smart grids: A multi-agent deep reinforcement learning framework. *arXiv* **2022**, arXiv:2211.15858. [CrossRef]
6. Yao, Z.; Lum, Y.; Johnston, A.; Mejia-Mendoza, L.M.; Zhou, X.; Wen, Y.; Aspuru-Guzik, A.; Sargent, E.H.; Seh, Z.W. Ml for a sustainable energy future. *Nat. Rev. Mater.* **2023**, *8*, 202–215. [CrossRef]
7. Nakabi, T.A.; Toivanen, P. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustain. Energy Grids Netw.* **2021**, *25*, 100413. [CrossRef]
8. Selim, A.; Mo, H.; Pota, H. Optimal scheduling of grid supply and batteries operation in residential building: Rules and learning approaches. In *Proceedings of the IEEE 5th Student Conference on Electric Machines and Systems (SCEMS)*; IEEE: New York, NY, USA, 2022; pp. 1–6.
9. Taboga, V.; Bellahsen, A.; Dagdougui, H. Deep reinforcement learning for peak load reduction in aggregated residential houses. In *Proceedings of the IEEE Power Energy Society General Meeting, Montreal, QC, Canada, 2–6 August 2020*; IEEE: New York, NY, USA, 2020; pp. 1–5.
10. Su, Y.; Zhang, T.; Xu, M.; Tan, M.; Zhang, Y.; Wang, R.; Wang, L. Rough knowledge enhanced dueling deep Q-network for household integrated demand response optimization. *Sustain. Cities Soc.* **2024**, *101*, 105065. [CrossRef]
11. Leitão, J.; Fonseca, C.M.; Gil, P.; Ribeiro, B.; Cardoso, A. A compressive receding horizon approach for smart home energy management. *IEEE Access* **2021**, *9*, 100407–100435. [CrossRef]
12. Vázquez-Canteli, J.R.; Nagy, Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Appl. Energy* **2019**, *235*, 1072–1089. [CrossRef]
13. Athanasiadis, C.L.; Papadopoulos, T.A.; Kryonidis, G.C.; Doukas, D.I. A holistic and personalized home energy management system with non-intrusive load monitoring. *IEEE Trans. Consum. Electron.* **2024**, *70*, 3935–3946. [CrossRef]

14. Kiasari, M.M.; Aly, H.H. Climate-adaptive residential demand response integration with power quality-aware distributed generation systems. *Electronics* **2025**, *14*, 3846. [CrossRef]
15. Rahman, M.M.; Hasan, M.M.; Suleymanov, Y.; Dadon, S.; Saha, S.; Suki, T.T. Energy purchase optimization for microgrid systems using deep-Q-learning. In *Proceedings of the IEEE International Conference on Clean Electrical Power*; IEEE: New York, NY, USA, 2025; pp. 1–5.
16. Kahraman, A.; Yang, G. Home energy management system based on deep reinforcement learning algorithms. In *Proceedings of the IEEE PES ISGT-Europe, Novi Sad, Serbia, 10–12 October 2022*; IEEE: New York, NY, USA, 2022; pp. 1–5.
17. Peirelinck, T.; Hermans, C.; Spiessens, F.; Deconinck, G. Combined peak reduction and self-consumption using proximal policy optimisation. *Energy AI* **2024**, *16*, 100323. [CrossRef]
18. Charbonnier, F.; Morstyn, T.; McCulloch, M.D. Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility. *Appl. Energy* **2022**, *314*, 118825. [CrossRef]
19. Lai, B.C.; Chiu, W.Y.; Tsai, Y.P. Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty. *IEEE Trans. Emerg. Topics Comput. Intell.* **2022**, *6*, 568–579. [CrossRef]
20. Bahrami, S.; Chen, Y.C.; Wong, V.W.S. Deep reinforcement learning for demand response in distribution networks. *IEEE Trans. Smart Grid* **2020**, *12*, 1496–1506. [CrossRef]
21. Ye, Y.; Qiu, D.; Wang, H.; Tang, Y.; Strbac, G. Real-time autonomous residential demand response management based on twin delayed deep deterministic policy gradient learning. *Energies* **2021**, *14*, 531. [CrossRef]
22. Cheng, L.; Huang, P.; Zhang, M.; Yang, R.; Wang, Y. Optimizing electricity markets through game-theoretical methods: Strategic and policy implications for power purchasing and generation enterprises. *Mathematics* **2025**, *13*, 373. [CrossRef]
23. Coccato, S.; Barhmi, K.; Lampropoulos, I.; Golroodbari, S.; van Sark, W. A review of battery energy storage optimization in the built environment. *Batteries* **2025**, *11*, 179. [CrossRef]
24. Saroha, P.; Singh, G.; Lilhore, U.K.; Simaiya, S.; Khan, M.; Alroobaea, R.; Alsafyani, M.; Alsufyani, H. Dynamic appliance scheduling and energy management in smart homes using adaptive reinforcement learning techniques. *Sci. Rep.* **2025**, *15*, 24594. [CrossRef]
25. Li, Y.; Zhang, X.; Gao, W.; Qiao, J.H. Lessons learnt from the residential zero carbon district demonstration project, governance practice, customer response, and zero-energy house operation in Japan. *Front. Energy Res.* **2022**, *10*, 915088. [CrossRef]
26. Ruelens, F.; Claessens, B.J.; Vandael, S.; Schutter, B.D.; Babuka, R.; Belmans, R. Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Trans. Smart Grid* **2018**, *8*, 2149–2159. [CrossRef]
27. Parisio, A.; Rikos, E.; Glielmo, L. A model predictive control approach to microgrid operation optimization. *IEEE Trans. Control Syst. Technol.* **2014**, *22*, 1813–1827. [CrossRef]
28. Xue, W.; Jia, N.; Zhao, M. Multi-agent deep reinforcement learning based HVAC control for multi-zone buildings considering zone-energy-allocation optimization. *Energy Build.* **2025**, *329*, 115241. [CrossRef]
29. Du, W.; Huang, X.; Zhu, Y.; Wang, L.; Deng, W. Deep reinforcement learning for adaptive frequency control of island microgrid considering control performance and economy. *Front. Energy Res.* **2024**, *12*, 1361869. [CrossRef]