energies

MDPI

Article

Reinforcement Learning for EV Fleet Smart Charging with On-Site Renewable Energy Sources

Handong Li [1,*], Xuewu Dai [2], Stephen Goldrick [1], Richard Kotter [2], Nauman Aslam [2] and Saleh Ali [3]

[1] Department of Biochemical Engineering, University College London, London WC1H 0AW, UK; s.goldrick@ucl.ac.uk
[2] Mathematics, Physics and Electrical Engineering, Northumbria Univerisity, Newcastle upon Tyne NE1 8ST, UK; xuewu.dai@northumbria.ac.uk (X.D.); richard.kotter@northumbria.ac.uk (R.K.); nauman.aslam@northumbria.ac.uk (N.A.)
[3] School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK; saleh.ali@ncl.ac.uk
* Correspondence: handong.li@ucl.ac.uk

Abstract: In 2020, the transportation sector was the second largest source of carbon emissions in the UK and in Newcastle upon Tyne, responsible for about 33% of total emissions. To support the UK's target of reaching net zero emissions by 2050, electric vehicles (EVs) are pivotal in advancing carbon-neutral road transportation. Optimal EV charging requires a better understanding of the unpredictable output from on-site renewable energy sources (ORES). This paper proposes an integrated EV fleet charging schedule with a proximal policy optimization method based on a framework for deep reinforcement learning. For the design of the reinforcement learning environment, mathematical models of wind and solar power generation are created. In addition, the multivariate Gaussian distributions derived from historical weather and EV fleet charging data are utilized to simulate weather and charging demand uncertainty in order to create large datasets for training the model. The optimization problem is expressed as a Markov decision process (MDP) with operational constraints. For training artificial neural networks (ANNs) through successive transition simulations, a proximal policy optimization (PPO) approach is devised. The optimization approach is deployed and evaluated on a real-world scenario comprised of council EV fleet charging data from Leicester, UK. The results show that due to the design of the rewards function and system limitations, the charging action is biased towards the time of day when renewable energy output is maximum (midday). The charging decision by reinforcement learning improves the utilization of renewable energy by 2–4% compared to the random charging policy and the priority charging policy. This study contributes to the reduction in battery charging and discharging, electricity sold to the grid to create benefits and the reduction in carbon emissions.

Keywords: reinforcement learning; renewable energy; EV fleet; ORES

check for updates

Citation: Li, H.; Dai, X.; Goldrick, S.; Kotter, R.; Aslam, N.; Ali, S. Reinforcement Learning for EV Fleet Smart Charging with On-Site Renewable Energy Sources. *Energies* **2024**, *17*, 5442. https://doi.org/10.3390/en17215442

Academic Editor: Branislav Hredzak

Received: 20 September 2024
Revised: 20 October 2024
Accepted: 28 October 2024
Published: 31 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Aligned with the UK's 2050 net-zero target, a core objective of the energy research theme is to enhance the efficient use of renewable energy through digital technologies. Charging electric vehicles (EVs) using renewable energy is a vital approach toward carbon-neutral road transport, which remains the second largest source of carbon emissions in the UK. The shift toward electrifying business fleets has recently become a significant strategy for lowering transportation-related CO_2 emissions [1]. Although the Global Covenant of Mayors (GCoM) notes that fleet electrification can reduce direct emissions, this alone does not guarantee a national or global reduction in emissions if EVs are charged with non-renewable electricity [2]. Increasing the integration of renewable energy sources (RESs) into EV charging is key; however, existing guidance lacks a clear focus on managing EV charging through renewable sources. While large-scale RES projects remain critical, the

trend is moving toward smaller, on-site renewable installations (ORES), which offer greater opportunities for EVs to utilize on-site renewable energy generation, potentially reducing overall emissions from road transport.

In recent years, a number of smart charging strategies for electric vehicle (EV) fleets have been developed, primarily focusing on optimizing charging schedules in response to grid constraints and energy prices. Rule-based approaches, such as those in [3], propose time-of-use strategies where charging occurs during off-peak hours. Optimization-based methods, including the work in [4], utilize mixed-integer programming or heuristic algorithms to minimize costs or emissions while respecting operational constraints. For instance, the study in [5] presents an approach that sizes and plans charging infrastructure for commercial EV fleets. Their method incorporates real-world operational data and highlights the importance of smart charge management (SCM) in optimizing infrastructure and reducing peak demands. Additionally, renewable energy sources (RESs) have increasingly been integrated into EV charging systems. The review by [6], thoroughly examines various smart charging strategies that incorporate RESs, energy storage systems, and smart grids to reduce carbon footprints. More recent studies have begun to explore reinforcement learning (RL) methods for EV fleet management, as seen in [7], where summarized RL algorithms are applied to dynamically optimize charging decisions under uncertainty. Despite these advancements, existing methods often overlook the combined uncertainties of weather patterns and charging demand. This paper addresses this gap by utilizing multivariate Gaussian distributions derived from historical data to model both weather-related and demand uncertainties, thus enabling more robust decision-making in the presence of on-site renewable energy sources (ORES).

The power generation from on-site renewable energy sources (ORES) is highly variable, leading to supply-side fluctuations. Temporary battery storage can help mitigate these fluctuations. On the demand side, EV charging needs are also uncertain due to varied usage patterns and tasks that create dynamic travel and charging requirements. Transitioning EV energy from reliable, grid-based fossil electricity to variable on-site ORES introduces a key challenge: designing a charging strategy that maximizes self-consumption of renewable energy while meeting unpredictable EV charging needs, minimizing energy costs, and reducing investment in storage with minimal grid impact. Studies show that as the share of ORES rises, large-capacity energy storage becomes essential, leading to higher electricity storage investments [8]. Vehicle-to-grid (V2G) technology has been suggested to reduce reliance on static storage; however, V2G contributes to battery degradation and requires intelligent charging strategies to manage or offset these additional costs [9]. In addition to the well-established mathematical programming methods (LIP [10], MILP [11], DP [12], and their variants [13]), reinforcement learning and AI have been recently proposed to maximize the self-consumption of ORES [14]. Recent studies take into consideration uncertainties either at the RES supply side or the EV demand side by Markov chain or time series modeling. However, the interplay of the stochastic characteristics of EV demands and ORES supplies should be considered carefully when developing and implementing the smart charging method for real-world problems. There is currently a shortage of study in this field, particularly on how to develop an empirically data-driven model, as opposed to the ideal theoretical model. Efforts have been made in the past to develop smart charging, but they have typically been based on an ideal theoretical model and have neglected to fully account for uncertainties in the demand and ORES supply for electric fleets. In the real world, ideal models are approximations based on empirical data, which poses difficulties for the majority of existing solutions.

Huang et al. addressed the challenge of formulating the stochastic matching problem as a Markov decision process (MDP), which they solved approximately using event-based optimization (EBO) to mitigate the curse of dimensionality [15]. Sadeghi-Barzani et al. introduced a Mixed-Integer Non-Linear Programming (MINLP) approach for the optimal placement and sizing of fast-charging stations [16]. The task of aligning the variable wind supply with EV charging demand to minimize costs is both complex and practically sig-

nificant, prompting Luo et al. to create a novel metric to evaluate grid impact without relying on full power flow calculations [17]. Jiao et al. developed hybrid strategies for energy generation and charging operations aimed at enhancing the environmental benefits of EV charging services integrated with nonrenewable energy sources in the context of photovoltaic (PV) station planning [18]. Dong et al. proposed a probability-based energy management strategy using a deep reinforcement learning model with continuous action space, effectively managing the time-dependent dynamics of EVs and the grid [19]. Long et al. parameterized the aggregated charging policy with the incomplete Beta function, capitalizing on problem structure [20]. In peer-to-peer (P2P) power trading, where renewable energy integration poses challenges due to the inconsistent production of wind and PV power, Lee et al. applied a reinforcement learning (RL) approach to balance resource production [21].

In the study of the reinforcement learning algorithm, this paper will focus on the proximal policy optimization (PPO) algorithm. PPO is a popular algorithm in the field of reinforcement learning, and it has been applied to a wide range of control and decision-making problems. In recent years, PPO has been used in the field of EV charging to optimize the charging process and improve the overall efficiency of the power grid. One of the main advantages of PPO is its ability to effectively balance exploration and exploitation. This is particularly important in the context of EV charging, where the charging process is influenced by a complex and dynamic set of factors, such as the availability of renewable energy sources, the state of the power grid, and the behavior of EV owners. PPO has been used to optimize the charging schedule and power levels of EV chargers in order to minimize the cost of electricity and ensure the power grid remains stable. Several studies have applied PPO to EV charging, each one with a different approach and goal. For example, in a study by Chen et al., PPO was used to optimize the charging schedule of EVs in a microgrid, with the goal of minimizing the cost of electricity while ensuring that the power grid remains stable. The authors showed that the PPO-based approach was able to achieve significant cost savings compared to traditional charging methods [22]. Li et al. used PPO to optimize the power levels of EV chargers in a smart grid, with the goal of reducing the peak load on the grid and improving the overall efficiency of the system [23]. It showed that the PPO-based approach was able to achieve significant improvements in the peak load and overall efficiency of the system compared to traditional methods. Zhang et al. used PPO to optimize the charging schedule of EVs in a smart grid, with the goal of reducing the overall cost of electricity while ensuring that the power grid remains stable. The paper showed that the PPO-based approach was able to achieve significant cost savings compared to traditional charging methods [24].

This paper aims to explore the potential for smartly integrating the dynamic charging needs of electric vehicle (EV) fleets with the fluctuating output from On-site Renewable Energy Sources (ORES) through a data-driven reinforcement learning (RL) decision support tool. The tool is designed to provide real-time, optimal dispatch and charging schedules for commercial or public authority EV fleets, maximizing ORES self-consumption despite uncertainties. The objective is to meet variable EV charging demands while minimizing energy costs, storage investments, and impact on the electricity grid. In the case study, both a depot/deployment location (main source) and public charging stations (secondary) are considered for a commercial organization or local council to achieve carbon-neutral road transport. This paper is a feasibility study of optimized real-time dispatching and charging management of business/council Electric Fleets (EF) using on-site renewable energy sources (ORES) and is designed to answer the open question of how to improve the efficient use of highly variable ORE while meeting the dynamic travel demands of a collaboration of EVs in business fleets at the lowest possible cost of facility investment and with the least impact on the grid. This is addressed by exploring the digital technology of data-driven artificial intelligence and optimization to manage and match the dynamic demands (dispatching and charging of electric fleets) and the supply of ORES—both of which are highly variable—in order

to increase ORES utilization and efficiency, thus reducing the actual emission of road transport. The novel technical idea is the development of an empirical data-driven intelligent decision support tool that can provide a tactical dispatching and charging schedule to improve the self-consumption of on-site renewable energy. Our objectives are to enhance the utilization of renewable energy, decrease reliance on the mains grid, and do so without increasing customer wait times. This feasibility study on optimized real-time dispatching and charging management of business/council Electric Fleets (EF) using ORES is designed to address the open question of how to improve the efficient use of highly variable ORES while meeting the dynamic travel demands of collaborative EVs in business fleets at the lowest possible cost of facility investment and least impact on the grid. The novel technical idea presented here is the development of an empirical data-driven intelligent decision support tool that can provide a tactical dispatching and charging schedule to improve the self-consumption of on-site renewable energy. In summary, the main contributions of this paper are as follows:

- Provide a more practical method for maximizing self-consumption of ORES through empirical data-driven optimal dispatching and scheduling of corporate and municipal Electric Fleets (EF) in an urban environment.
- Using a multivariate Gaussian distribution to generate enormous amounts of data for the agent to learn, a simulation model of reinforcement learning was built.
- Regarding the construction of the model, this research does not discuss the reuse of batteries. In reality, this can significantly reduce the loss of batteries due to charging and discharging and maximize the self-consumption of ORES, hence boosting the system's efficiency and minimizing carbon emissions.
- Employing data from a real Leicester EF, this research verifies the capability of algorithmic decision-making using AI decisions that can improve overall renewable energy utilization by 2–4% in the extreme case study shown.

The structure of the paper is as follows: Section 2 proposes an optimal control formulation for the charging schedule with the fleet composition modeled as a Markov decision process. Section 3 presents an environment setup based on deep reinforcement learning. In Section 4, the PPO algorithm's configuration and algorithmic flow are shown. In addition, the optimal design model is implemented, and the arrangement of an electric vehicle (EV) charging station in Leicester is evaluated through a case study. Section 5 concludes with some findings and recommendations for future research.

2. Problem Hypothesis

This paper will develop a more feasible solution to the maximization of self-consumption of ORES by the empirical data-driven optimal dispatching and scheduling of business and municipal Electric Fleets (EF) in an urban area. The EFs consist of a set of heterogeneous EVs with various dynamic travel demands, which may change from day-to-day operations. As shown in Figure 1, the main EV recharging facility is at the fleet's depot, assisted by several public charging stations in an urban area as the secondary charging facility. The primary charging facility at the depot includes multiple ORES (such as PV and wind generators), a static battery energy storage system (BESS), and a range of EV chargers (AC or DC). These components are linked to a 400 V local distribution network, which connects to the main grid through a transformer. The Energy Hub unit, consisting of unidirectional or bidirectional inverters, links the ORES and EV charging stations, managing energy flow between the supply and demand sides.

In order to transform this practical problem into a reinforcement learning simulation, this study will make the following assumptions. (1) We simply evaluate energy transformation and do not account for energy losses caused during energy transformation. (2) The renewable energy inputs are wind and solar power, both of which are mathematically modeled for energy creation, as will be detailed in greater detail below. (3) This paper does not address battery recharging for EV fleets, hence optimizing the use of renewable energy.

We consider that this system (Figure 1) consists of a set of N EVs as $\mathcal{N} = \{1, 2, \ldots, N\}$. Each vehicle $n \in \mathcal{N}$ refers to the vehicle ID.

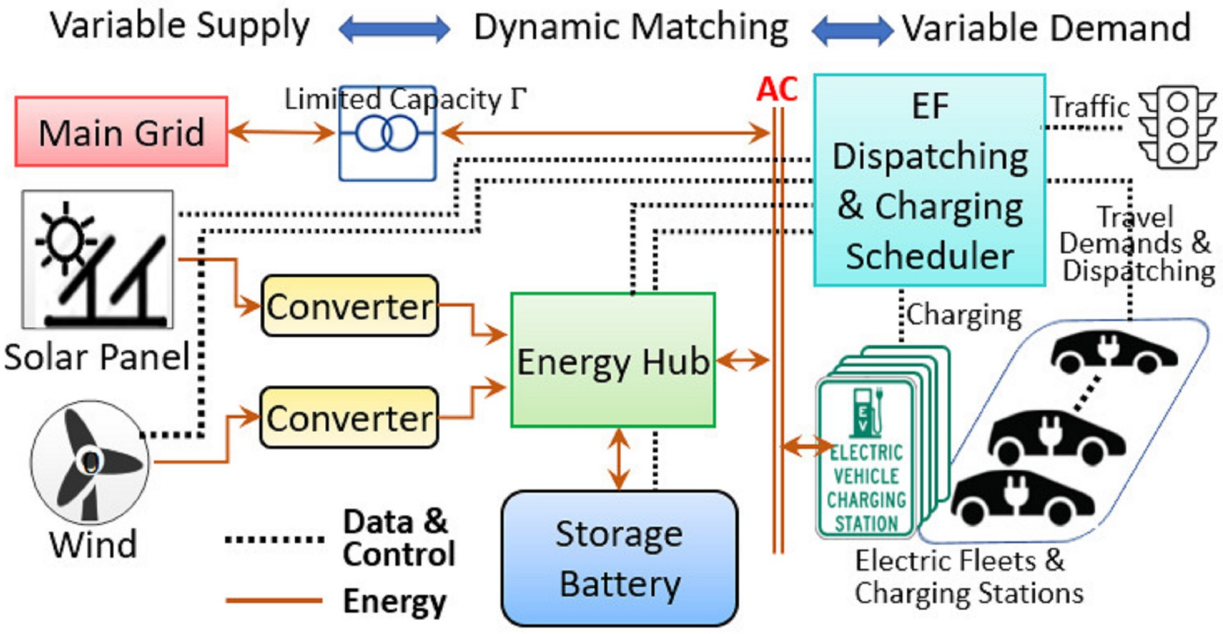


Figure 1. Diagram of the self-generated renewable energy for the electric fleet.

The optimization problem is written as a Markov decision process (MDP), with the charging system represented by the state vector s_k over decision stage k. Therefore, this paper designs the system state function s_k as:

$$s_k = [ETBC_k, EG_k, T_k, TP_k] \tag{1}$$

where $ETBC_k$ is the energy required at the time k. $ETBC_k = \sum_{i=i}^{n} E_k^i$, E_k^i is the single vehicle energy required at time k, and n is the total number of vehicles. TP_k is the remaining parking time. T_k is the current time period depending on the time interval. EG_k is the renewable energy generation from the wind speed and solar radiance. The decisions to be taken for time $k + 1$ are represented by the decision vector a_k^n = [0,1], where 1 indicates charging the car and 0 indicates no charging from time k to time $k + 1$, and n is the vehicle identification number. Given the state-decision pair (a_k^n, s_k) at stage k, the MDP follows the dynamic transition $s_{k+1} \sim \mathcal{P}(a_k^n, s_k)$ from stage k to stage $k + 1$ and earns a reward r_{k+1} from the reward function $\mathcal{R}(a_k^n, s_k)$.

2.1. Transition Dynamics

The following describes the transition dynamics $s_{k+1} \sim \mathcal{P}(a_k^n, s_k)$ over decision stages k = 0, 1,..., K. Given a_k^n and s_k, we first determine the charging power P_v to charge the vehicles to the time k+1 in s_{k+1} based on the action a_k^n as:

$$ETBC_{k+1} = ETBC_k - 0.25 \cdot \sum_{i=1}^{n} a_k^i \cdot P_v \tag{2}$$

where the value of 0.25 is derived from the system's resolution. This paper's system resolution is 15 min; hence, the day will be separated into 96 time periods. Therefore, the value of K is 95.

2.2. Reward Signal Design

The reward signal r_{k+1} linked with the transition process $s_{k+1} \sim \mathcal{P}(a_k^n, s_k)$ is now formulated. When developing the incentive signal for the charging EVs scheduling problem, we consider two objectives. From the standpoint of the charging station operator, we strive to reduce the amount of electricity consumed from the main grid. From the user's perspective, we would try to reduce their waiting costs, which may conflict with the charging station's interests. The reward signal r_{k+1} for charging EVs $k + 1$ is, therefore, expressed as:

$$r_{k+1} = -EM_k - \alpha EN_k \tag{3}$$

which is a linear combination of the energy borrowed from the main grid, EM_k, and the not fully charged energy, EN_k. The minus sign in Equation (3) indicates that we are attempting to maximize rewards, as they represent expenses in the present situation. The trade-off parameter α is pivotal in determining the balance between the two objectives. When $\alpha = 1$, it signifies that the interests of the charging station and the users are regarded equally in the optimization settings. This flexibility allows us to prioritize one objective over the other depending on specific operational needs. For instance, if we want to emphasize reducing customer wait times, we can increase α. Conversely, if the goal is to maximize renewable energy utilization, a lower α would be more appropriate. This design satisfies our requirements, as we anticipate that there will be a substantial amount of clean energy available, and we want to ensure that vehicles do not leave without meeting their charging demands. The details of the energy from the main grid EM_k and the not fully charged energy EN_k are provided in the following sections.

2.2.1. Energy from the Main Grid

The cost of recharging the fleet of electric vehicles at k + 1 is comprised of both renewable energy and energy from the main grid. The purpose of this study is to utilize as much renewable energy as possible. Consequently, energy from the main grid can be detrimental as,

$$EM_k = EG_k - ETBC_k \tag{4}$$

where in order to meet demand when there is insufficient renewable energy, the energy must be supplemented from the main grid.

2.2.2. Not Fully Charged Energy

In the beginning, decision-making on the reinforcement learning is uncertain, and we are concerned about the reinforcement agent's excessive use of renewable energy, which could result in a situation where only a small amount of renewable energy is generated and no charging action is being applied during the parking period. As a result, the reward signal is designed as follows to accommodate for this circumstance.

$$EN_k = \sum_{i=1}^{n} \sum_{k=k_s}^{k} (E_k^i - 0.25 \cdot a_k^i \cdot P_v) \tag{5}$$

where k is the time at present, and k_s is the time at which the car begins parking. This equation indicates how much of the required power remains unsatisfied until now. This allows the EV to be charged as rapidly as possible while assuring the usage of renewable energy.

The purpose of the MDP model, given the definitions of EM_k and EN_k, is to maximize the number of rewards collected throughout all the transitions (1 day for which $K = 96$), i.e.,

$$\max_{a_k^i} R = \sum_{k=1}^{K} r_k \tag{6}$$

2.2.3. Constraints

Operational limitations are imposed on the maximization Equation (6). First, we have the limits for the system variables: SOC_k and TP_k for each service run k:

$$C^{min} \leq E_k^{SOC} \leq C^{max}, \tag{7}$$

$$0 \leq TP_k \leq TP^{max}, \tag{8}$$

$$E_k^{SOC}, TP_k \in \mathbb{Z}^+. \tag{9}$$

where $\mathbb{Z}^+$ represents the set of positive numbers. C_{min} and C_{max} are the minimum and maximum capacities of the battery, taking into account safety, the size of the battery, and the minimum capacity. E_k^{SOC} is a cumulative charging variable that represents the battery's charge at time k. Each charge should be smaller than the design's maximum charge capacity. Consequently, the SOC_k should be $SOC_k = (E_k^{SOC} + C^{min})/C^{max}$. Similarly, TP_k specifies the maximum parking duration. While this model does not explicitly address detailed EV battery degradation, the focus on maximizing the direct utilization of renewable energy helps mitigate the frequency of battery charging and discharging cycles. This strategy not only conserves battery health by reducing charge/discharge events but also lowers the required battery capacity, as more energy can be supplied directly from renewable sources. As a result, the overall volume and cost of batteries can be reduced, making the system more efficient and sustainable in the long term.

3. Environment Model

3.1. Modeling with Power Generation Uncertainties

3.1.1. Wind Power Model

The electrical output of wind turbines is generated by converting the mechanical power derived from wind energy via the turbine blades. According to aerodynamic principles, the mechanical power generated by a wind turbine, as stated in [25], is given by:

$$P_m = \frac{1}{2}\rho A u^3(t) C_p(\theta, \lambda) \tag{10}$$

where ρ is the air density, P_m represents the mechanical power from the wind turbine, A is the blade-swept area, and u denotes wind speed. $C_p(\theta, \lambda)$ indicates the proportion of available wind energy captured. θ is the pitch angle, and λ is the tip speed ratio. Assuming a constant rotational speed, the electrical power output can be simplified as:

$$P_m = \frac{1}{2}\rho A u^3(t) \tag{11}$$

This model uses a 10 kW horizontal axis wind turbine (HAWT), selected for its common application in small- to medium-scale renewable installations, making it representative of on-site renewable energy sources (ORES). The technical specifications for this 10 kW HAWT, based on [26], are summarized in Table 1, where u_r represents the rated wind speed, u_c represents the cut-in speed the minimum wind speed for power generation), and u_f is the furling speed, where the turbine halts to prevent damage. P_{er} is the 10 kW rated power. The power output behavior assumes a rise between u_c and u_r and then remains constant between u_r and u_f, and dropping to zero outside these ranges. This relationship can be expressed as:

$$P_m(t) = \begin{cases} 0 & , \text{if } u(t) \leq u_c \\ \frac{P_{er}(u_c^3 - u^3(t))}{u_c^3 - u_r^3} & , \text{if } u_c < u(t) \leq u_r \\ Per & , \text{if } u_r < u(t) \leq u_f \\ 0 & , \text{if } u_f \leq u(t) \end{cases} \tag{12}$$

Thus, Equation (3) allows wind energy generation to be evaluated based on expected wind speed and the wind turbine specifications.

Table 1. Specifications of the 10 kW wind turbine generator [26].

Parameter	Values
Type	Three blade upwind
Rated power (kW): P_{er}	10
Start-up wind speed (m/s): u_s	3.4
Cut in wind speed (m/s): u_c	3.1
Rated wind speed (m/s): u_r	13
Furling wind speed (m/s): u_f	15.6
Max. design wind speed (m/s): u_m	54
Rotor shaft speed (rpm): R	60–350

3.1.2. PV Model

For the solar power model, we adopt the formulation from Akhbari et al. [27], where the DC power output of a photovoltaic (PV) system is expressed as a function of the effective solar irradiance, system efficiency, and surface area of the PV panels. This model is chosen due to its simplicity and effectiveness in capturing the key factors that influence solar power generation, particularly irradiance, which varies with location, time, and weather conditions. The model incorporates temperature effects on PV efficiency, allowing for a more accurate representation of the energy output across different environmental conditions. Akhbari et al. expressed the DC power provided by a solar PV source as:

$$P_{dc}(t) = I_{eff}(t)\eta_g A_s \tag{13}$$

In this model, $I_{eff}(t)$ represents the incident efficient radiance (Wm^{-2}), A_s stands for the effective surface area of the solar PV source (m^2) and η_g denotes the efficiency of the solar PV source. As described in [28], the efficiency of solar PV sources depends on factors such as ambient temperature, temperature loss coefficient, and nominal operating cell temperature. Since solar power output is linearly related to the incident radiance $I_{eff}(t)$ in this model, accurately estimating $I_{eff}(t)$ is crucial for forecasting solar energy production. The PV model parameters used in the simulation is from the solar system at Northumbria University. Hence, the parameters for the 30 kW solar PV system are shown in Table 2.

Table 2. Parameters for the 30 kW solar power system [29].

Parameter	Values
Peak power (kW)	29.84
Production (estimate) (kWh/year)	23,024.00
Panels	Seraphim SPR-6PB-265, Changzhou, China
Area (m^2): A_s	188.00
Inverter	Solar Edge SE17000 and SE4000, Tel Aviv, Israel
Orientation	180° (SOUTH)
Slope	11°

3.2. Weather Tensor Components Multivariate Normal Distribution

A multivariate Gaussian distribution will be employed in order to create a massive amount of environmental data for the AI algorithm training. In order to increase the interaction between AI and the environment, this research assumes that both the weather data and the parking demand data conform to the Gaussian distribution. In the weather simulation, it contains seven meteorological variables that must be simulated in the environment. which are temperature ($T_e(t)$, °), sunshine duration ($S_{Dur}(t)$, minutes), solar radiance ($I_{eff}(t)$, $\mathrm{W}_{\mathrm{m}}^{-2}$), cloud cover total ($C_C(t)$, %), mean sea level pressure ($P_{sea}(t)$, hPa),

wind speed ($u(t)$, m/s), and wind direction ($W_D(t)$, degrees). This part organizes the collected meteorological data and defines the random meteorological variables as follows:

$$X_w(t) = \left[T_e(t), S_{Dur}(t), I_{eff}(t), P_{sea}(t), C_C(t), u(t), W_D(t)\right] \tag{14}$$

At this stage, it follows from the definition of the multivariate normal distribution that the multivariate normal distribution holds for a p-dimensional random vector X. The density is represented as:

$$f(X) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right], X \in \mathbb{R}^p \tag{15}$$

Then, X follows a p-dimensional normal distribution, and x is a p-dimensional normal variable in which $X \sim N_p(\mu, \sigma)$. μ is the mean matrix of X. Σ is the covariance matrix of X and a symmetric positive definite (SPD) matrix. The mean value of the random vector X component X_i is $E(X_i) = \mu_i, i = 1, 2, 3, \ldots, p$. Thus, the mean vector is:

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \mu \tag{16}$$

where μ is a p-dimensional column vector, called the mean vector. The covariance matrix of X is:

$$\begin{aligned} D(X) &= E[(X - E(X))(X - E(X))^T] \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}_{p \times p} \stackrel{def}{=} \Sigma \end{aligned} \tag{17}$$

where

$$\sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))], i, j = 1, 2, \cdots, p \tag{18}$$

Through the multiplication of matrices and vectors, it can be determined that $(X-\mu)^T\Sigma^{-1}(X-\mu)$ is a number. This is the Mahalanobis distance between the vectors X and μ. When Σ equals I, it represents the Euclidean distance between the two vectors. This is a simulation of energy-related meteorological data. In addition, a Gaussian multivariate distribution will also be used to simulate data for the EV charging demand. In this paper, the charging demand for the EV fleets will use 3 variables as follows:

$$X_D(t) = \left[T_p(t), P_{com}(t), E_{req}(t)\right] \tag{19}$$

where $T_p(t)$ is the parking time or connecting time. We assume the vehicles can be charged while parked. At each of these times, the vehicle is parked for a distinct amount of time. The parking time is typically shorter at lunchtime and longer during the evening. $P_{com}(t)$ is the probability that an EV is coming to park for charging purposes.

4. Proposed Reinforcement Learning Algorithm with Proximal Policy Optimization

This section describes an on-policy reinforcement learning (RL) solution strategy for the MDP. First, we provide the construction of the policy function employed in the RL framework with operational constraints incorporated. The authors then provide a solution framework based on the proximal policy optimization (PPO) method.

PPO is particularly well-suited to our problem due to the continuous nature of the action space (charging decisions) and the need for computational efficiency in handling large datasets derived from wind, solar, and EV demand simulations. Additionally, PPO has

demonstrated strong performance in environments with high variability and uncertainty, such as the one presented in our study, where both renewable energy generation and EV charging demands are stochastic. By employing PPO, we ensure that the learning process is robust and can adapt to these uncertainties, providing a reliable charging strategy for the EV fleet. The PPO algorithm is a policy-based reinforcement learning algorithm that uses two neural networks. By feeding the current state of the agent into the neural network, the action and reward are eventually obtained. The state of the agent is then updated according to the action, and the objective function containing the reward and action is used to apply a ladder based on the objective function containing the reward and action. The weight parameters in the neural network are updated using gradient ascent so that an action judgment that results in a greater overall reward value can be obtained.

4.1. Policy Function

Then, we establish a differentiable stochastic policy function π with parameter θ (sometimes denoted as π_θ) such that $\pi(x|s) \in [0,1]$ represents the probabilistic distribution. The probability of taking action a in state s_k under the current policy is given by:

$$\pi(a|s) = \sigma(f(s_k, a_k)) \tag{20}$$

where $f(s_k, a_k)$ is the output of the policy neural network, which is a function that maps states and actions to a scalar value. The sigmoid function, $\sigma(x)$, is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{21}$$

The sigmoid function maps the output of the policy neural network in the range [0, 1], which represents the probability of the action being taken. The value function, which estimates the expected return from state s_k under the current policy, is given by:

$$V_\phi(s_k) = E[G|S = s_k] \approx v_\phi(s_k) \tag{22}$$

where G is the return, which is the sum of the discounted rewards received after entering the state, and $v(s_k)$ is the output of the value function neural network, which is a function that maps states to a scalar value. The value function is used to evaluate the expected return of following the current policy from a given state.

The advantage function, which estimates the difference between the expected return of an action taken in state s_k and the expected return of the action with the highest expected return in that state, is given by:

$$A(s_k, a_k) = Q(s_k, a_k) - V(s_k) \approx A(s_k, a_k) \tag{23}$$

where $Q(s_k, a_k)$ is the action-value function, which estimates the expected return of taking a specific action in a specific state, and $A(s_k, a_k)$ is the output of the advantage function neural network, which is a function that maps states and actions to a scalar value. The advantage function is used to evaluate the relative goodness of an action taken in a given state compared to other actions that could be taken in that state.

The ratio of the new policy to the old policy is given by:

$$r(\theta) = \frac{\pi(a|s;\theta)}{\pi(a_K|s_k;\theta_{old})} \tag{24}$$

where θ_{old} represents the parameters of the old policy, and θ represents the parameters of the new policy. The ratio is used to compare the probability of taking an action under the new policy to the probability of taking that action under the old policy.

The surrogate loss function, which is used to compute the gradient of the objective function with respect to the policy parameters, is given by:

$$L = E[min(r(\theta)A, clip(r(\theta), 1-\epsilon, 1+\epsilon)A)] \tag{25}$$

where $\epsilon \in (0,1)$ is a fixed parameter whose upper and lower bounds for $r(\theta)$ are defined by the interval $[1-\epsilon, 1+\epsilon]$.

This loss function is an approximation of the objective function and is used to update the policy parameters. The min function serves to prevent the loss from becoming too far from the previous policy, while the clip function ensures that the loss is well-behaved by preventing the ratio from becoming too large. Therefore, choosing a *theta* that maximizes L^e would result in the selection of decisions that yield greater rewards in subsequent rounds. In accordance with [30], we employ a generalized advantage estimation (GAE) where,

$$A_k = \delta_k + (\gamma\lambda)\delta_{k+1} + \ldots + (\gamma\lambda)^{K-k+1}\delta_{K-1} \tag{26}$$

where the discount factor $\gamma \in [0,1]$ and the GAE parameter $\lambda \in [0,1]$, balancing the estimation biases and variances. The term δ_k in Equation (26) is referred to as a one-step temporal difference (TD) residual as in [31,32].

$$\delta_k = \begin{cases} r_{k+1} - V_\phi(s_k), & if\ k = K-1 \\ r_{k+1} + \gamma V_\phi(s_{k+1}) - V_\phi(s_k), & otherwise \end{cases} \tag{27}$$

where $r_{k+1} + \gamma V_\phi(s_{k+1})$ represents the value of making a decision, a_k is equal to the sum of the current reward r_k and the VFA's anticipated future reward $\gamma V_\phi(s_{k+1})$ [31]. The parameters ϕ of $V_\phi(s_k)$ must be iteratively updated by minimizing a mean squared error function as

$$\phi_{i+1} = arg\ \min_{\phi} L^e(\phi) \tag{28}$$

The PPO technique is an on-policy RL algorithm that optimizes the policy function with ANN parameters using the policy gradient method [33] and the trust-region method [34]. Additionally, PPO employs a VFA V_ϕ parameterized by an ANN with parameters. Unlike the policy ANN, which is used to make decisions based on the current state s_k at each stage k, the VFA $V_\phi(s_k)$ is used to estimate the predicted total reward that may be earned from stage k and beyond by being in state s_k at the current stage k. Algorithm 1 is a classic PPO algorithm that summarizes the preceding equations and the concepts in this work.

Algorithm 1 Proximal Policy Optimization(PPO) Algorithm

```
Input: initialize algorithm policy parameter θ and value function parameter φ_0, set time step t = 0
for e = 1, 2, ..., E (e is the episode) do
    for k = 1, 2, ..., 96 (k is the daily 15-min time interval) do
        Collect data by interacting with environment
        Estimate value function v(s) and advantage function A(s, a)
        Compute ratio r(θ) = π(a|s; θ) / π(a|s; θ_old)
        Optimize policy by minimizing L using gradient descent
        Set θ_old = θ
        Increment time step t
    end for
end for
```

4.2. Implementation Process of Proposed PPO Method

The flowchart in Figure 2 demonstrates the implementation of our proposed method for solving the EV fleet charging allocation scheduling problem. Specifically, the algorithm initializes the power rate, the dissatisfaction factor, the discount rate, and the learning rate

at the starting time slot, i.e., k = 1. At each time interval, the algorithm internally predicts the output of both solar panels, and the wind turbine generates the renewable energy for the next time interval (15 min) (Equations (12) and (13)). After obtaining the predicted weather information, the PPO algorithm is used to make the ideal energy scheduling decision for the charging time of the EV with different IDs. In each episode e, the agent examines the current state s_k and may take a random action. Immediately following the execution of an action, the agent obtains the reward $r(s_k, a_k)$, observes the subsequent state s_{k+1}, and updates it. The value of L is updated by Equations (24) and (25). This process is repeated until the end of the episode. After an event, the agent checks the termination criterion, i.e., $|L^e - L^{e-1}| \leq \tau$, where τ is a system-dependent parameter that controls the accuracy of convergence. If this termination requirement is not met, the agent will proceed to the next episode and repeat the steps outlined in the previous section. Finally, each agent will obtain the best action for each time interval, i.e., k = 1, 2, . . . 96. In this technique, 96 time intervals of one day are considered as one episode. Note that only the best hourly action is available. The above procedure will be continued until the final hour, or until k = 96. Finally, the cumulative daily award value is determined. When this value converges, i.e., when training concludes, this model can be applied to real data for evaluation.

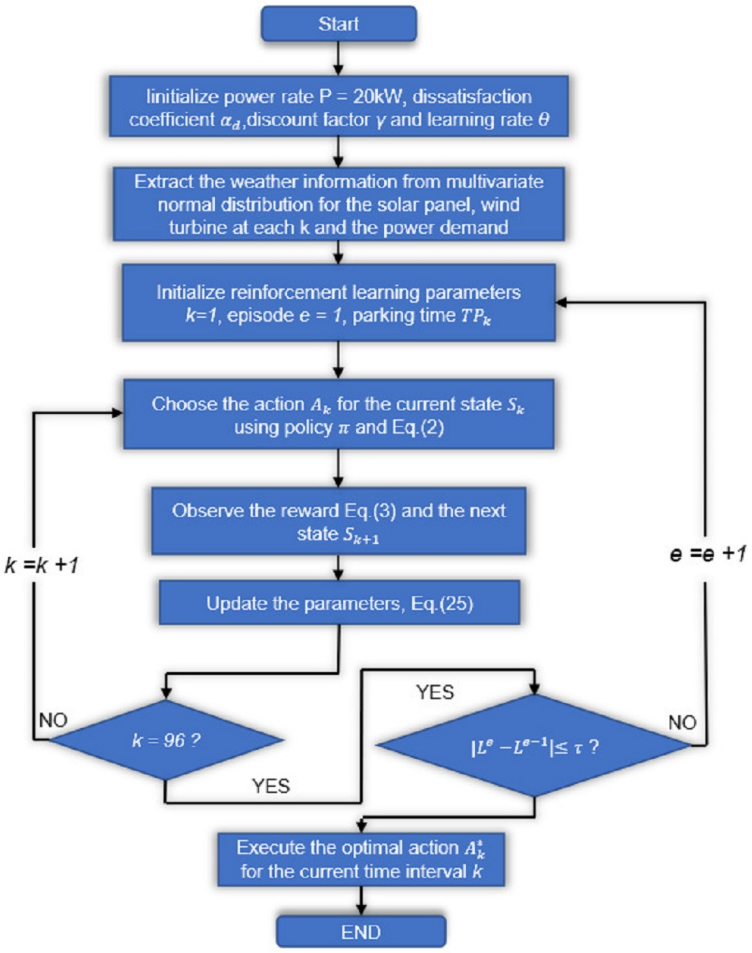


Figure 2. Implementation of our proposed solution flowchart.

The data used in the paper are from a real network in Leicester, as shown in Table 3. This information is related to four EVs (Nissan Leaf), each EV is assigned a specific number (12BA, 9490, b4e7 and 840A), and the charge information of each is collected with a radio frequency identification (RFID) charging card. Leicester City Council Planning, Development, and Transportation, as well as energy departments, have provided mea-

surements/available data from the Leicester City Hall operational pilot baseline. EV users can connect to any charge point at a specified hour of the day, disconnect their connection at another specified hour, and exchange electricity with the network during the period between the connection and disconnection. Information on how EV users charge their vehicles was collected from the period (3 April 2018, to the 29 March 2019) in Table 1. Each row of Table 1 indicates a one-time connection and disconnection of one of the EVs to a charge point during the year, which shows 1217 charging behaviors of EV users. These charging behaviors include information about start time, stop time, duration of connection to charger, amount of energy exchanged with the network and RFID related to EVs.

Table 3. Vehicle charging data stored.

Label	Start Time	Stop Time	Charging Duration	Energy (kWh)	RFID
1	3 April 2018 10:10	3 April 2018 11:09	00:59:06	3.48	****13BA
2	3 April 2018 11:36	5 April 2018 08:41	1.21:05:02	3.09	****B4E7
3	3 April 2018 12:13	5 April 2018 10:09	1.21:56:30	9.88	****840A
...					
1217	29 March 2019 14:59	1 April 2019 09:04	2.18:04:43	3.91	****13BA

**** The data pertains to the username and has been concealed to protect privacy.

4.3. Power Demand Side

This study compiled the data from the table above into the probability of parked vehicles for these four IDs per 24 h, as shown in Figure 3. As seen in this figure, the EV fleet is typically parked within the depot most of the time. The city council only employs these vehicles around midday to accomplish certain jobs so that they have enough time to recharge. This condition is advantageous for the design since we must choose the most cost-effective and environmentally beneficial time to recharge during these hours. Since we have defined each episode as a day, this study is not concerned with the analysis of nighttime cars but rather with the scheduling of daytime charging. In addition, the parking probability of the B4E7 is substantially lower than that of the other vehicles due to the fact that it undergoes repairs and component updates at a set time, but we have retained this variation to add complexity to the data.

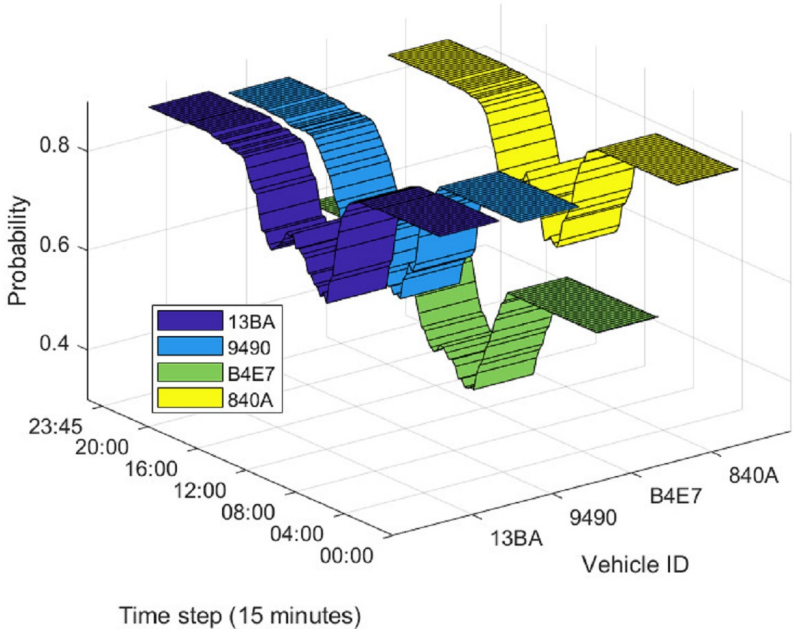


Figure 3. Parking probability graph for an average of 24 h of 4 vehicles.

In conjunction with Figures 4 and 5, it is possible to statistically determine the maximum, minimum, variance, and mean values of the parking probability, energy demand, and parking duration for each ID vehicle in each 15-min interval. Using these data and the multidimensional Gaussian distribution described in Section IIIb, a charging demand model for the EV fleet is created.

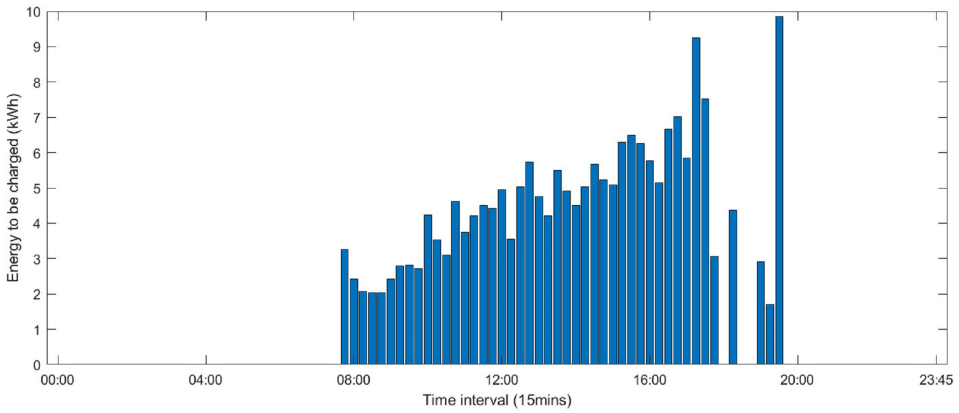


Figure 4. Energy requirement for the EV fleet.

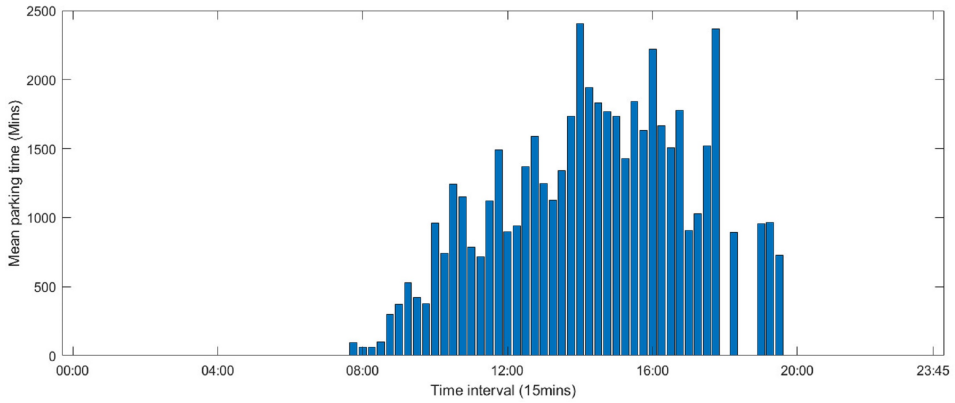


Figure 5. Mean parking time for the EV fleet.

4.4. Renewable Energy Generation Simulation

This paper compares typical UK statistics for the same time of year as the charging data for EV fleets [35]. Among these are temperature ($T_e(t)$, °), sunshine duration ($S_{Dur}(t)$, minutes), solar radiance ($I_{eff}(t)$, W_m^{-2}), cloud cover total ($C_C(t)$, %), mean sea level pressure ($P_{sea}(t)$, hPa), wind speed ($u(t)$, m/s), wind direction ($W_D(t)$, degrees). As the temporal resolution of this paper is 15 min, the aforementioned data will be updated to ($T_e(k)$, $S_{Dur}(k)$, $I_{eff}(k)$, $C_C(k)$, $P_{sea}(k)$, $u(k)$, $W_D(k)$) based on the time interval. Since the resolution of the meteorological data presented in this research is in hours. This section of the simulation thus assumes that hourly weather variations are uniform. Therefore, the weather information is the same for four 15-min intervals every hour. As in the real world, the algorithm trained the model to determine by itself how much energy it would generate over the following interval. This paper does not directly model the relationship between weather data and energy generation but provides the agent with historical weather information as observations. The agent is informed of the prior day's weather information and all relevant weather data for the current day from time intervals 1 to k. While we use a multivariate Gaussian distribution to generate weather data, this is to train the AI with more data, as real-world data are limited. Though real weather patterns are more complex, this method allows for efficient training, and future work will explore more detailed models to capture greater accuracy.

In this study, the models in Tables 1 and 2 use the weather data in order to calculate Figures 6 and 7's 24-h average solar and wind energy production. As indicated by the figure, the peak generation of renewable energy occurs around noon. In summary, as explained in section IIIb, this study can generate a multidimensional Gaussian distribution using the maximum, minimum, mean, and variance of the meteorological data. This simulation can be used to model the energy generation and the weather model in order to provide training data for the reinforcement learning algorithm.

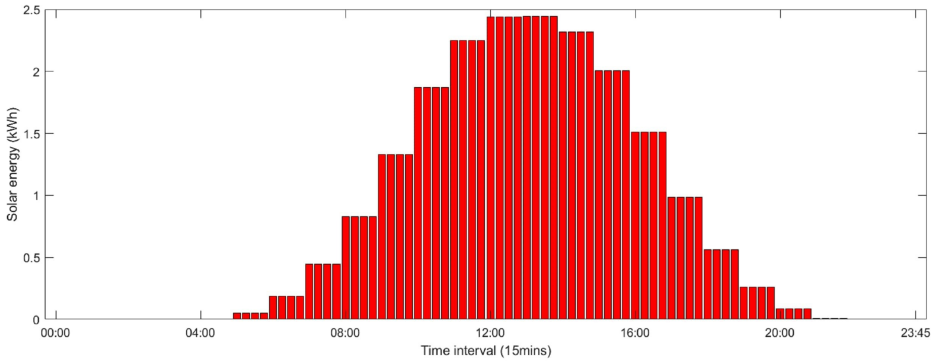


Figure 6. Solar power generation.

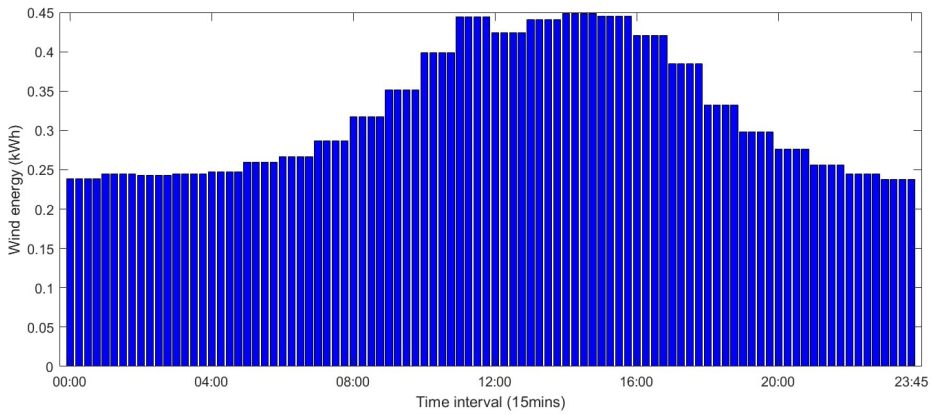


Figure 7. Wind power generation.

4.5. Simulation Result

Figure 8 shows the results of the PPO algorithm's training. As explained in the previous design, this research uses the PPO algorithm to train the agent to predict the weather for the next time interval. As illustrated in the figure, the accumulated reward acquired from the RL method based on the PPO algorithm tends to stabilize with more episodes as the built-in prediction accuracy increases and converges on the optimal solution. The figure represents the oscillation of the learning curve, which is caused by the stochastic policy function's exploration of the state and decision space. All learning curves converge after around 20,000 training episodes, demonstrating that the cumulative reward score climbs steadily over the rounds and stabilizes later in the training process, with all learning curves converging beyond this point.

In order to evaluate the effectiveness of the proposed algorithm, this study presents three charging strategy scenarios as benchmarks for comparison. The selected dataset reflects real-world conditions, which may lead to comparatively low energy consumption during the year analyzed. Notably, the charging times align closely with peak renewable energy generation, particularly around midday. This statistical insight suggests that while the algorithm is designed to enhance renewable energy utilization, its impact on overall consumption may not be significantly pronounced.

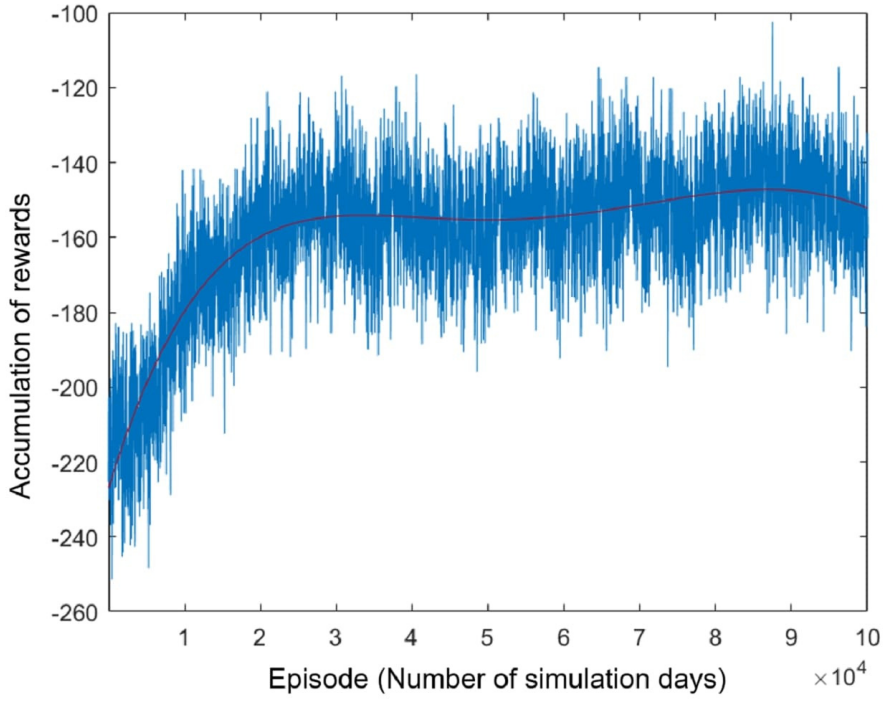


Figure 8. PPO algorithm training iterations.

The study begins by establishing a random charging strategy that allows the EV fleet to charge at any time while parked. This serves as a baseline scenario. Second, a priority charging strategy is implemented, which aligns with typical charging behavior, initiating charging as soon as the vehicle begins to park. Lastly, we introduce the Proximal Policy Optimization (PPO) algorithm, which optimizes charging schedules to prioritize renewable energy availability, effectively functioning as a renewable energy priority charging strategy. By comparing these three approaches, we aim to demonstrate the advantages of the PPO algorithm in maximizing renewable energy utilization while accommodating the real-world complexities of EV charging behavior.

Figure 9 shows the proportion of electricity taken from the mains grid for the three charging strategies, 33%, 31%, and 29% for the random charging approach, the priority charging strategy, and the reinforcement learning charging strategy, respectively. Observably, the random charging strategy has the lowest utilization of renewable energy. The random charging consumes a large percentage of the nighttime hours, and as shown in Figures 6 and 7, the rest of the time period generates the least amount of renewable energy; thus, a significant amount of energy will be taken from the main grid. The time-priority charging approach utilizes 69% renewable energy. In conclusion, the reinforcement learning strategy suggested in this paper will boost utilization to 71%. It is important to highlight that this is not the limit of renewable energy usage. This inaccuracy is a result of the algorithm's inability to anticipate the next weather conditions with perfect accuracy. It still includes the uncertainty of weather information.

Although the model's predictions are not entirely free of uncertainty, a significant advantage is that the AI has been trained using the distribution of weather data for an entire year. This ensures that the model is well-acquainted with various weather patterns and uncertainties, providing it with the robustness to handle most real-world meteorological variability. By leveraging this comprehensive dataset, the reinforcement learning algorithm is able to adjust charging times to maximize renewable energy utilization. While minor inaccuracies may still arise due to the assumption of uniform weather within hourly intervals, the model has demonstrated strong performance in aligning charging times with peak energy generation. Future enhancements could focus on incorporating even more precise short-term weather forecasts to further reduce reliance on grid electricity.

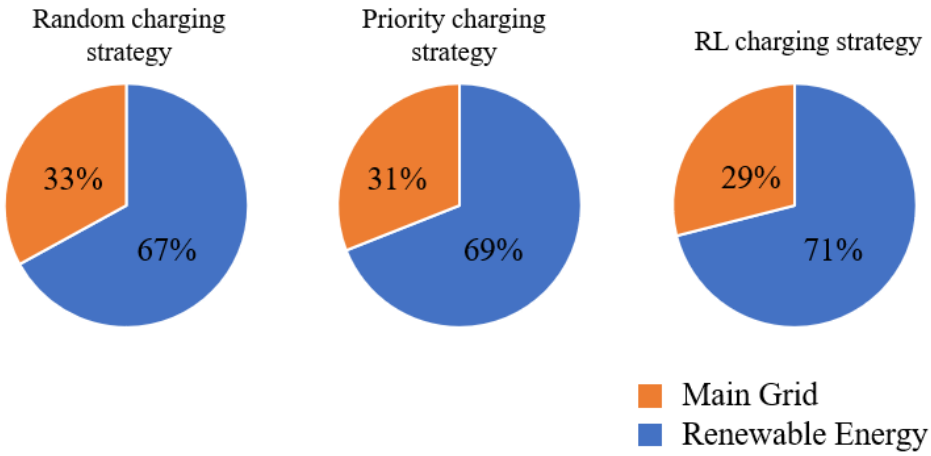


Figure 9. Comparison of three different charging strategies.

4.6. Validation

To study how reinforcement learning makes EV fleet charge decisions, we selected the data of the Leicester council depot on 28 March 2021 and the typical weather data for that date in the UK. As illustrated in Figure 10, two vehicles are to be charged on the selected day. At 12:30, vehicle 1 and vehicle 2 present their respective energy needs for the red and blue energy columns. The green pins illustrate the distribution of renewable energy generation according to the time of day. It is worth mentioning that this distribution is not shown to the agent. Therefore, the agent can only predict future energy output based on the meteorological circumstances, as previously presented.

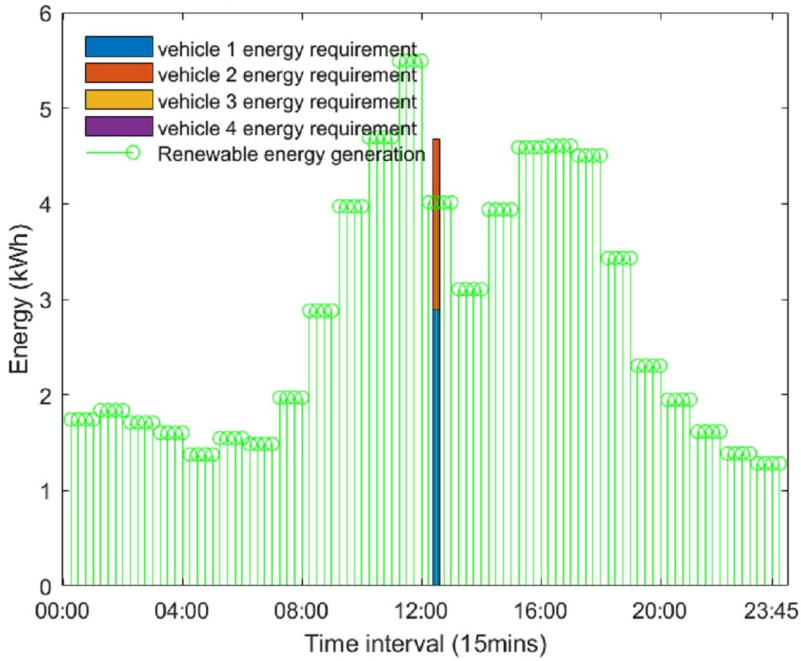


Figure 10. Renewable energy generation and the energy requirement distribution.

As demonstrated in Figure 11, according to the time-priority charging policy, charging of the EVs begins at 12:45. The figure illustrates that the energy required to charge the two energy volumes of vehicles is significantly more than the energy generated by wind and solar. It can also cover the energy requirements for this charge in 15 min at the charging rate proposed in this paper. Consequently, the system will request the extra electricity from the main grid.

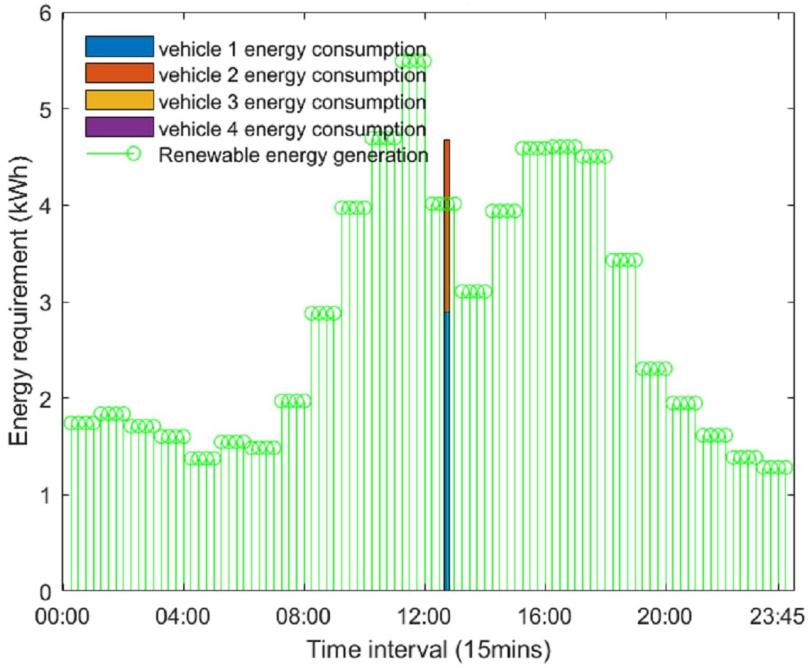


Figure 11. Priority charging strategy.

The charging plan that was developed with the assistance of the AI, which predicts that there will not be adequate generation of renewable energy in the subsequent time period, is shown in Figure 12. In conclusion, this approach consumes an additional 15 min of time, but it does not require power from the main grid, which results in an increase in the overall economy of the EV fleet as well as a decrease in carbon emissions. Hence, the priority strategy requires 0.65 kWh of energy from the main grid. In contrast, the AI does not require energy from the main grid.

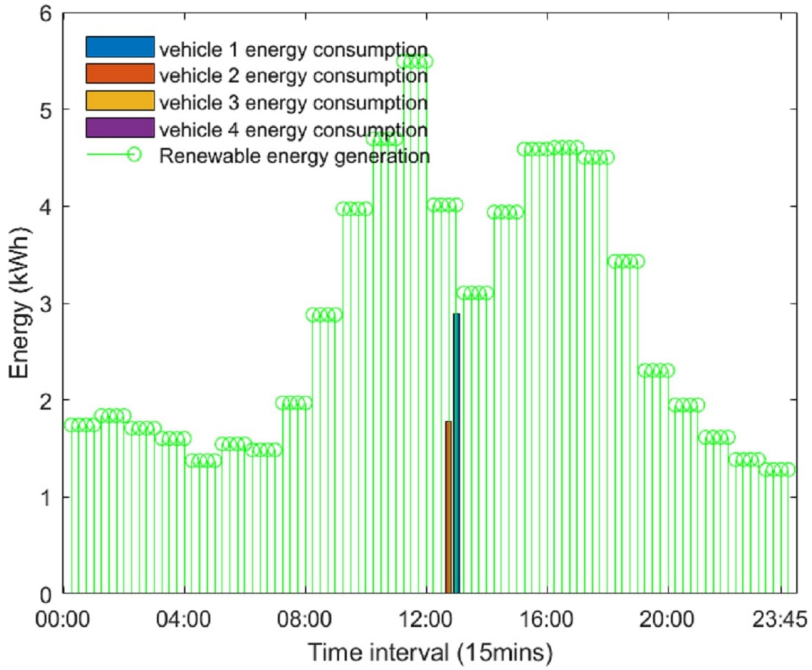


Figure 12. Reinforcement learning charging strategy.

5. Conclusions

In conclusion, this paper proposed a data-driven method for optimizing the use of ORES in an urban environment by using an RL algorithm called PPO. By utilizing real-world data, such as typical UK weather data and data from a Leicester council EV charging depot, and employing the PPO algorithm, our method was able to predict short-term renewable energy generation and charging demand for electric vehicle fleets. The simulation results demonstrate that the proposed PPO RL method can effectively reduce electricity costs and carbon emissions while improving computational efficiency. As per the test results, our method improved the overall renewable energy utilization by 2–4% in the case study. This research makes a valuable contribution to the field of ORES optimization and EV fleet management, highlighting the potential of AI-based decision-making in improving overall renewable energy utilization. However, it is important to note that the proposed method was tested using a specific case study, and further research is needed to generalize the findings to other urban environments and EV fleet sizes. Future work may also consider incorporating other factors such as EV battery degradation and user behavior into the optimization model.

While the proposed PPO-based reinforcement learning method shows promising results in optimizing renewable energy utilization for EV fleet charging, there are several real-world deployment challenges to consider. First, the model relies on accurate weather forecasts to predict renewable energy generation, and any forecasting errors could lead to suboptimal charging decisions. Although the model has been trained using a full year's weather data to account for a wide range of uncertainties, extreme weather events or unforeseen shifts in weather patterns may still pose challenges in real-world applications. Additionally, the method assumes the availability of high-quality historical weather and charging data, which may not be accessible in all urban environments. Another potential challenge is the scalability of the model. While the case study demonstrated its effectiveness for a specific fleet size, deploying the model across larger EV fleets with diverse usage patterns may require further adjustments to maintain computational efficiency and accuracy. Finally, real-world deployment would need to consider factors such as EV user behavior, charging station availability, and infrastructure limitations, which could influence the model's performance in practical scenarios.

Author Contributions: Conceptualization, H.L. and X.D.; methodology, N.A.; software, H.L.; validation, H.L., S.G. and S.A.; formal analysis, R.K.; investigation, R.K.; resources, X.D.; data curation, H.L. and X.D.; writing—original draft preparation, H.L.; writing—review and editing, all authors; visualization, N.A.; supervision, X.D. and S.G.; project administration, X.D.; funding acquisition, X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the EPSRC project Electric Fleets with On-site Renewable Energy Sources (EFORES) under grant EP/W028727/1, the Wuhan AI Innovation Program (2022010702040056), and the EU Interreg North Sea Region programme's SEEV4-City (Smart, clean Energy and Electric Vehicles for the City) project (J-No.: 38-2-23-15).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request and the algorithm of this research can be found in https://github.com/handongli2019/Reinforcement-learning-for-Microgrid-management (accessed on 1 October 2024).

Acknowledgments: The authors would like to acknowledge Adrian McLoughlin from Newcastle City Council and James Yu from Scottish Power for their support with the application and technical discussions.

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

References

1. Li, H.; Wu, H.; Gulati, I.; Ali, S.A.; Pickert, V.; Dlay, S. An improved sliding mode control (SMC) approach for enhancement of communication delay in vehicle platoon system. *IET Intell. Transp. Syst.* **2022**, *16*, 958–970. [CrossRef]
2. Bergen, E.; Hoogt, J.; Warmerdam, J.; Herteleer, B.; Rimmer, C.; Kotter, R.; Bentley, E. *Vehicle 4 Energy Services (V4ES) Evaluation for Upscaling and Transnational Potential: Assessing the Potential of Further Roll-Out of 8 Differing V4(ES) Solutions*; Interreg: Lille, France, 2020; p. 1.
3. Suyono, H.; Rahman, M.T.; Mokhlis, H.; Othman, M.; Illias, H.A.; Mohamad, H. Optimal scheduling of plug-in electric vehicle charging including time-of-use tariff to minimize cost and system stress. *Energies* **2019**, *12*, 1500. [CrossRef]
4. Neri, M.; Guelpa, E.; Verda, V. Design and connection optimization of a district cooling network: Mixed integer programming and heuristic approach. *Appl. Energy* **2022**, *306*, 117994. [CrossRef]
5. Mohamed, A.A.; Ingalalli, A.; Rehman, W.U.; Bhavaraju, V. Sizing of Fast Charging Infrastructure Site for Commercial Electric Vehicle Fleet Considering Smart Charge Management. In Proceedings of the 2024 IEEE Transportation Electrification Conference and Expo (ITEC), Anaheim, CA, USA, 18–20 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
6. Barman, P.; Dutta, L.; Bordoloi, S.; Kalita, A.; Buragohain, P.; Bharali, S.; Azzopardi, B. Renewable energy integration with electric vehicle technology: A review of the existing smart charging approaches. *Renew. Sustain. Energy Rev.* **2023**, *183*, 113518. [CrossRef]
7. Abdullah, H.M.; Gastli, A.; Ben-Brahim, L. Reinforcement learning based EV charging management systems—A review. *IEEE Access* **2021**, *9*, 41506–41531. [CrossRef]
8. Pamparana, G.; Kracht, W.; Haas, J.; Díaz-Ferrán, G.; Palma-Behnke, R.; Román, R. Integrating photovoltaic solar energy and a battery energy storage system to operate a semi-autogenous grinding mill. *J. Clean. Prod.* **2017**, *165*, 273–280. [CrossRef]
9. Taibi, E.; del Valle, C.F.; Howells, M. Strategies for solar and wind integration by leveraging flexibility from electric vehicles: The Barbados case study. *Energy* **2018**, *164*, 65–78. [CrossRef]
10. Jahannoosh, M.; Nowdeh, S.A.; Naderipour, A.; Kamyab, H.; Davoudkhani, I.F.; Klemeš, J.J. New hybrid meta-heuristic algorithm for reliable and cost-effective designing of photovoltaic/wind/fuel cell energy system considering load interruption probability. *J. Clean. Prod.* **2021**, *278*, 123406. [CrossRef]
11. Alberizzi, J.C.; Rossi, M.; Renzi, M. A MILP algorithm for the optimal sizing of an off-grid hybrid renewable energy system in South Tyrol. *Energy Rep.* **2020**, *6*, 21–26. [CrossRef]
12. Wei, Q.; Shi, G.; Song, R.; Liu, Y. Adaptive dynamic programming-based optimal control scheme for energy storage systems with solar renewable energy. *IEEE Trans. Ind. Electron.* **2017**, *64*, 5468–5478. [CrossRef]
13. Kartite, J.; Cherkaoui, M. Study of the different structures of hybrid systems in renewable energies: A review. *Energy Procedia* **2019**, *157*, 323–330. [CrossRef]
14. Lissa, P.; Deane, C.; Schukat, M.; Seri, F.; Keane, M.; Barrett, E. Deep reinforcement learning for home energy management system control. *Energy AI* **2021**, *3*, 100043. [CrossRef]
15. Huang, Q.; Jia, Q.S.; Xia, L.; Guan, X. Event-based optimization for stochastic matching EV charging load with uncertain renewable energy. In Proceedings of the 11th World Congress on Intelligent Control and Automation, Shenyang, China, 29 June–4 July 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 794–799.
16. Sadeghi-Barzani, P.; Rajabi-Ghahnavieh, A.; Kazemi-Karegar, H. Optimal fast charging station placing and sizing. *Appl. Energy* **2014**, *125*, 289–299. [CrossRef]
17. Luo, C.; Huang, Y.F.; Gupta, V. Stochastic dynamic pricing for EV charging stations with renewable integration and energy storage. *IEEE Trans. Smart Grid* **2017**, *9*, 1494–1505. [CrossRef]
18. Jiao, Z.; Lu, M.; Ran, L.; Shen, Z.J.M. Infrastructure Planning of Photovoltaic Charging Stations. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3560677 (accessed on 2 October 2024).
19. Dong, C.; Sun, J.; Wu, F.; Jia, H. Probability-based energy reinforced management of electric vehicle aggregation in the electrical grid frequency regulation. *IEEE Access* **2020**, *8*, 110598–110610. [CrossRef]
20. Long, T.; Jia, Q.S.; Wang, G.; Yang, Y. Efficient real-time EV charging scheduling via ordinal optimization. *IEEE Trans. Smart Grid* **2021**, *12*, 4029–4038. [CrossRef]
21. Lee, S.; Jin, H.; Nengroo, S.H.; Heo, T.; Doh, Y.; Lee, C.; Har, D. Reinforcement Learning Based Cooperative P2P Energy Trading between DC Nanogrid Clusters with Wind and PV Energy Resources. *arXiv* **2022**, arXiv:2209.07744.
22. Chen, X.; Li, Y.; Xu, L.; Liu, H. Optimizing electric vehicle charging schedules in microgrids using proximal policy optimization. *IEEE Trans. Smart Grid* **2020**, *11*, 2571–2580.
23. Li, Y.; Chen, X.; Liu, H.; Xu, L. Optimizing electric vehicle charging power in smart grids using proximal policy optimization. *IEEE Trans. Smart Grid* **2019**, *10*, 3015–3025.
24. Zhang, T.; Ding, Y.; Li, Y.; Wang, L. Proximal Policy Optimization based Electric Vehicle Charging Schedule Optimization in Smart Grid. *IEEE Trans. Smart Grid* **2021**, *12*, 947–957.
25. Luo, H.; Hu, Z.; Zhang, H.; Chen, H. Coordinated active power control strategy for deloaded wind turbines to improve regulation performance in AGC. *IEEE Trans. Power Syst.* **2018**, *34*, 98–108. [CrossRef]
26. Bowen, A.; Zakay, N.; Ives, R. The field performance of a remote 10 kW wind turbine. *Renew. Energy* **2003**, *28*, 13–33. [CrossRef]
27. Hossain, M.S.; Jahid, A.; Islam, K.Z.; Rahman, M.F. Solar PV and Biomass Resources-Based Sustainable Energy Supply for Off-Grid Cellular Base Stations. *IEEE Access* **2020**, *8*, 53817–53840. [CrossRef]

28. Bhattacharjee, A.; Samanta, H.; Ghosh, A.; Mallick, T.K.; Sengupta, S.; Saha, H. Optimized Integration of Hybrid Renewable Sources with Long-Life Battery Energy Storage in Microgrids for Peak Power Shaving and Demand Side Management under Different Tariff Scenario. *Energy Technol.* **2021**, *9*, 2100199. [CrossRef]
29. Power, A.R. Photovoltaic Production Monitoring. Available online: https://arpower.pvmeter.com/solar/App.html (accessed on 4 August 2024).
30. Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv* **2015**, arXiv:1506.02438.
31. Ying, C.S.; Chow, A.H.; Wang, Y.H.; Chin, K.S. Adaptive metro service schedule and train composition with a proximal policy optimization approach based on deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6895–6906. [CrossRef]
32. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
33. Daskalakis, C.; Foster, D.J.; Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5527–5540.
34. Kuba, J.G.; Chen, R.; Wen, M.; Wen, Y.; Sun, F.; Wang, J.; Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv* **2021**, arXiv:2109.11251.
35. Meteoblue. Hourly Historical Weather Simulation Data. Available online: https://www.meteoblue.com/en/historyplus (accessed on 30 May 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.