energies

MDPI

Article

Priority-Aware Sequential Decision-Making Reinforcement Learning in Energy Communities

Eduardo Gomes [1,2,3], Augusto Esteves [1,2], Hugo Morais [2,3,*] and Lucas Pereira [1,2]

[1] Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal; eduardo.c.gomes@tecnico.ulisboa.pt (E.G.); augusto.esteves@tecnico.ulisboa.pt (A.E.); lucas.pereira@tecnico.ulisboa.pt (L.P.)
[2] ITI, LARSyS, Técnico Lisboa, 1049-001 Lisboa, Portugal
[3] Instituto de Engenharia de Sistemas e Computadores—Investigação e Desenvolvimento, 1000-029 Lisboa, Portugal
* Correspondence: hugo.morais@tecnico.ulisboa.pt

Abstract

Renewable Energy Communities provide an innovative approach for decentralised energy management and cost reduction. However, coordinating diverse resources, particularly electric vehicles and storage systems, presents significant challenges, as simultaneous use often leads to operational conflicts. To address this, this work presents a sequential Multi-Agent Reinforcement Learning strategy in which agents execute actions in a specific order, leveraging the most recent environmental state information. Two existing priority-sorting mechanisms are utilised to verify the proposed approach: a fixed-priority mechanism and an adaptive approach informed by agents' contributions relative to the community. The models were trained on one month of data and evaluated against a simultaneous baseline using a two-year simulation. Results demonstrate that the sequential approach yields consistent cost savings, achieving up to a 7% monthly decrease and an overall 4% reduction compared to the simultaneous baseline.

Keywords: electric vehicles; Energy Communities; Multi-Agent Reinforcement Learning; renewable energy; sequential decision-making

1. Introduction

Driven by the global need to address climate change and carbon emissions, research on energy transition toward Renewable Energy Sources (RESs) becomes increasingly important, directly involving key issues such as energy efficiency and resource management [1]. The European Union has set a target of net-zero carbon emissions by 2050, further emphasising the need to explore the potential of the energy transition toward greener energy sources [2]. RES such as solar and wind are some of the most abundant, with solar being the most readily available and becoming increasingly cheaper over recent years [3].

Energy Communities (ECs) have emerged as a promising alternative to centralised energy production and consumption, empowering prosumers, reducing energy costs, and improving access to clean energy [4]. Examples such as Citizen Energy Communities (CECs) and Renewable Energy Communities (RECs) represent different implementations of an ECs, differing in participant criteria and resource types used [5]. Despite potential benefits, ECs face significant technical and social challenges such as integration of RESs and the uncertainty associated. Additional challenges include data availability, privacy concerns

Check for updates

Academic Editor: Valery Vodovozov

Received: 17 August 2026
Revised: 21 October 2026
Accepted: 1 November 2026
Published: 5 November 2026

Copyright: © 2026 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.

associated with information sharing [6], and the management of distributed RESs while balancing operational costs and member comfort [7].

Scalable, adaptive decision-making systems capable of operating under dynamic and potentially non-stationary conditions are in demand due to the previously mentioned challenges. While traditional optimisation or control methods can provide clear reasoning and optimality guarantees, these can struggle when rapid decision-making is required, or are exposed to settings where parties have access to partial information [8]. Multi-Agent Reinforcement Learning (MARL) offers a data-driven alternative that learns policies by having agents interact with the environment. However, the Reinforcement Learning (RL) and MARL approaches encountered in ECs settings assume simultaneous agent decision-making, potentially leading to coordination issues due to outdated information or conflicting actions [9]. An alternative to the simultaneous decision-making process is the sequential decision-making process, where agents are scheduled according to a ranking or priority system. Considering this gap, this work proposes a novel sequential MARL approach for managing ECs, drawing inspiration from the traditional system management approaches where different resources have varying priorities. By enforcing sequential execution with access to up-to-date environment states, the proposed approach minimises the risk of conflicting actions and promotes more efficient resource management. The proposed approach and a baseline MARL approach are compared using operational costs to validate the method. Finally, an adaptive priority-sorting mechanism is used in tandem with the sequential MARL approach to demonstrate the efficacy and flexibility of the proposed method and to provide an example of how it can be further expanded.

2. Related Work

Management of ECs and EC-related components has been extensively studied, utilizing methods such as Mixed-Integer Linear Programming (MILP), Model Predictive Control (MPC), game-theoretic formulations, and genetic algorithms [10–13]. These approaches offer optimality guarantees and transparency in decision-making, but can struggle with dynamic, decentralised, non-stationary environments.

Considering these limitations, there has been research on the application of other Machine Learning methods, specifically, RL in the context of ECs [14,15]. RL is a Machine Learning (ML) paradigm that can perform adaptive control and decision-making, focused on learning policies that maximise cumulative rewards [16]. MARL is an extension of RL to settings involving multiple decision-making agents, either cooperative or competitive, making it a well-positioned method for ECs, where agents must coordinate under potentially shared constraints.

Applications of MARL extend to tasks such as Electric Vehicle (EV) management [17,18], where the dynamic nature of EV charging and discharging must be taken into account. Recent literature has demonstrated the potential of Aviation-to-Grid applications, showcasing the application of Vehicle-to-Grid coupled with MARL in other domains [19], and multi-stage control frameworks for coordinating EV platoons [20]. In [21], MARL is applied to handle uncertainty in resource management, while other works focus on storage control [8], and Peer-to-Peer (P2P) energy trading [22–24]. For instance, ref. [22] utilised a fuzzy Q-Learning approach to optimise the energy trading process, increasing participant savings. Other MARL applications, such as energy pricing [25], handling grid stability [26], and blockchain-based MARL to control distributed resources in microgrids [27], can also be found in the literature.

Hierarchical approaches are also found in the literature, in which decisions are made at multiple levels. In [28], a hierarchical strategy is used for energy scheduling, highlighting the flexibility of Hierarchical Reinforcement Learning (HRL) in managing energy systems.

Other hierarchical methods focus on energy market settings [14,29], with some extending their scope to inter-community energy exchanges [30,31]. These approaches showcase the potential of HRL to handle decision-making processes formulated over multiple layers.

Sequential MARL mechanisms are part of recent literature, emerging as an alternative to simultaneous decision-making [32]. In [32], the authors introduce a sequential advantage function that attempts to credit agent participation by evaluating agents against each other, leading to improved coordination in StarCraft II. Similarly, sequential action rollouts are proposed in [33], showing improved learning efficiency without compromising convergence. The authors of [34] introduce Agent-by-Agent Policy Optimisation, where agents are sequentially updated while maintaining convergence improvement guarantees and outperforming conventional baselines. In [35], the authors use a sequential decision-making MARL framework to enhance the safety of autonomous vehicles and prioritise critical decisions based on the previous agent behaviours, thereby reducing traffic collisions and yielding higher overall rewards. Similarly, the authors of [36] introduce a sequential decision-making process across varying agent levels, conditioning later agents based on the behaviour of previous agents, exhibiting better performance in complex scenarios.

Sequential decision-making approaches based on priority mechanisms have been explored in the energy domain [37–39]. For example, Bilakanti et al. [37] introduce variable prioritisation coefficients to minimise renewable energy curtailment, while Ruan et al. [38] propose a sequential energy management method that improves user satisfaction in residential settings. Mussadiq et al. [39] evaluate the impact of three priority allocation strategies using Binary Integer Programming (BIP), showing gains in community energy savings.

These studies, though limited in number, collectively point to the promise of sequential decision-making in improving energy management outcomes. However, to the best of our knowledge, no existing work integrates sequential decision-making and MARL within the context of ECs. This represents a clear research gap. This paper addresses that gap by investigating whether a sequential MARL approach can feasibly and effectively reduce operational costs in ECs. Our results not only demonstrate the approach's potential, but also reveal avenues for future improvement.

3. Proposed Method: Sequential MARL for Energy Communities

This section provides information on agent behaviour, with Section 3.1 presenting the differences between the baseline MARL and the proposed sequential MARL formulation. Section 3.2 details the calculation of rewards and penalties for all agent groups used and provides agent observation and action spaces. Section 3.3 addresses the priority calculations utilised, presenting a simple ordering mechanism and an adaptive mechanism.

3.1. MARL and Sequential MARL

The most common approach for EC optimisation utilises a MARL formulation, in which resources are shared among agents. Each agent independently receives observations at the beginning of each timestep and determines its next action without knowledge of other agents' decisions. Such a lack of coordination can result in resource conflicts—for instance, agents might simultaneously attempt to draw energy from an already depleted source, such as a used storage unit, or unintentionally overuse resources.

Additionally, operating without up-to-date information can lead to redundant actions, such as multiple storage units selecting a discharge action to meet demand, which can result in excessive discharge and wasted resources. Figure 1 illustrates how independent agents interact with the environment without awareness of other agents' actions.

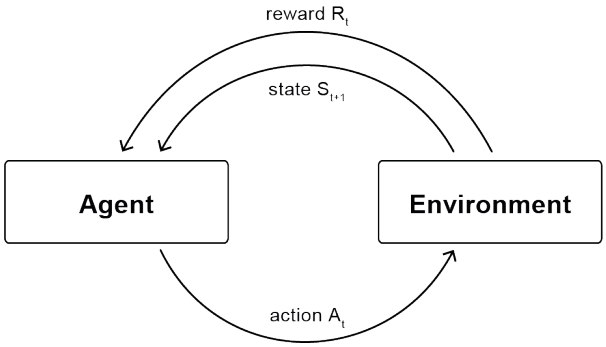


Figure 1. MARL agent and environment interactions.

Sequential approaches have been proposed in the literature to address these limitations [35,36]. In this formulation, agents act sequentially with access to updated environment states to inform their decision-making. This approach addresses potential wasteful behaviours, such as resource conflicts identified in the simultaneous baseline approach, and promotes agent coordination. Figure 2 illustrates how agents in the proposed sequential MARL approach act according to a predefined sequence and allow each agent to respond to an updated environment state, resulting from the effects of previous agents' actions. More concretely, agent m_0 receives an observation corresponding to the initial environment state (s_{m_0}), executes the corresponding action (a_{m_0}), and receives a reward (r_{m_0}). The following agent will then receive its observation based on the state of the environment (s_{m_1}), which is affected by the previous agent's action (a_{m_0}), and receive the corresponding reward. This process continues until the last agent executes its action and receives its reward. Furthermore, the sequential MARL approach allows defining priority sequences according to the community's needs and goals. This contrasts with Figure 3, where all the agents interact with the environment at the same time.

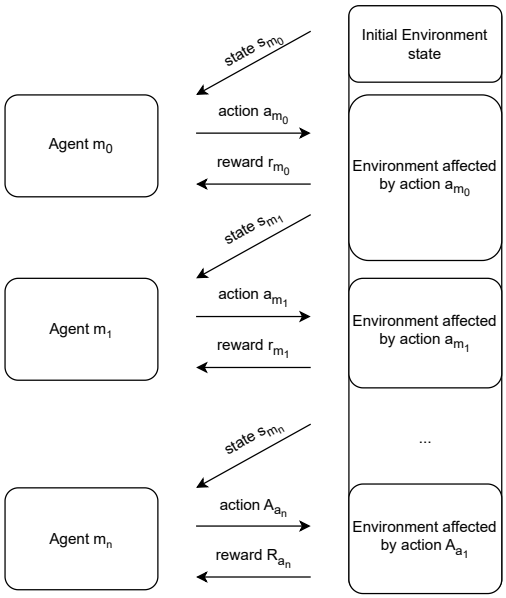


Figure 2. Agent–environment interactions under sequential MARL settings.

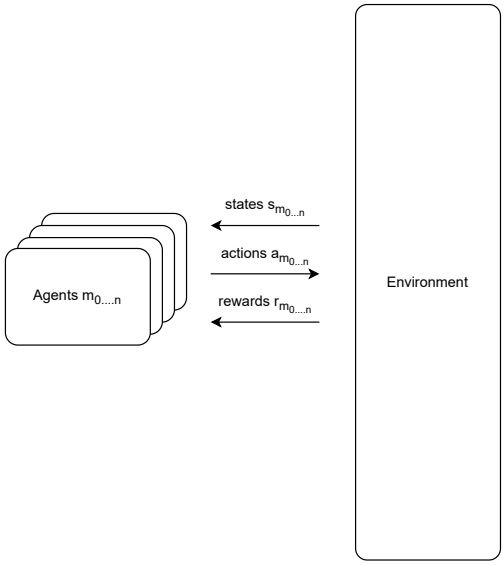


Figure 3. Agent–environment interactions under simultaneous MARL settings.

3.2. Policies, Costs, and Penalties

The MARL approaches used throughout this work categorise resources into three groups: Battery Energy Storage Systems (BESSs), EVs, and the main grid connector. Each agent within a group follows a customised policy, but all agents in the same group share a common observation and action space structure. For instance, all EVs observe the same set of variables, such as State of Charge (SoC), available energy, and grid prices, though the values of some of these variables may differ for each agent (i.e., each EV has its own SoC and trip schedule). Similarly, all agents in a group have the same set of actions. However, while the structure is shared, the outcomes of decisions are agent-specific. For example, an action to discharge 40% of an EV's maximum discharge value would produce different results depending on the vehicle characteristics. Policy sharing was initially considered but ultimately discarded as variability in resource characteristics made agents too heterogeneous for a shared policy to be viable.

The goal of the RL agents is to minimise the community's operational costs. As such, it can be expressed by Equation (1) as:

$$\min \sum_{t=0} \left(\phi_{ec,t} + \rho_{ec,t} + \sum_{i \in \mathcal{A}} (\phi_{i,t} + \rho_{i,t} + \phi_{ec,t} + \rho_{ec,t}) \right) \tag{1}$$

where $\phi_{ec,t}$ represents the cost of the community, $\rho_{ec,t}$ represents the penalty given to the grid connector, $\mathcal{A}$ denotes the set of agents, and $\phi_{i,t}$ and $\rho_{i,t}$ represent the cost and penalty of individual agent i at time step t, respectively. To better suit the RL framework, the objective is converted to the maximisation of the negative costs and penalties, as described by Equation (2):

$$\max \left(- \sum_{t=0} \left(\phi_{ec,t} + \rho_{ec,t} + \sum_{i \in \mathcal{A}} (\phi_{i,t} + \rho_{i,t} + \phi_{ec,t} + \rho_{ec,t}) \right) \right) \tag{2}$$

where all variables remained as defined for Equation (1).

Each agent's reward is given by their costs, received following an action, and the respective penalty according to agent-specific goals. Additionally, after the community manager balances the grid, the import/export costs are also given to every agent to incentivise community cooperation. This results in two reward attributions per timestep, once when the agent itself acts and once at the end of the timestep. This component is always the last to act, whether the simultaneous or sequential approaches are considered, isolated from the remaining agents, and functioning as a safety mechanism to ensure the community does

not suffer from excess or deficit of power. The following subsections provide additional detail on each resource group's specific cost and penalty calculations.

3.2.1. Battery Energy Storage Systems

In the studied community, the role of BESSs is to absorb excess energy generated beyond the required demand to meet the loads. BESS agents should discharge to assist the community in meeting its demand and charge whenever there is a surplus of energy from production. The observation space, as shown in Table 1, includes information on physical characteristics such as SoC, energy balance, grid pricing, and forecast community consumption and production. The bounds presented in Table 1 were defined to ensure the stability of the RL environment, with SoC being constrained to an interval of $[0, 1]$, representing the battery's stored energy as a percentage of its maximum capacity. Import and export prices are bound to the same range as SoC, with training prices ranging from a few cents to a dozen cents. Broader limits are used for other continuous variables to prevent observation space errors and handle unforeseen values. The latter are considered operational constraints that ensure the environment can safely transition to new states.

Table 1. BESS Observation Space. Composed of the BESS's characteristics, information on the energy pool, grid pricing, and forecasted community production and consumption.

Variable	Type	Bounds	Shape
State of Charge	float	$[0, 1]$	1
Capacity Max	float	$[0, 99{,}999]$	1
Available Energy	float	$[-99{,}999, 99{,}999]$	1
Maximum Charge	float	$[0, 99{,}999]$	1
Maximum Discharge	float	$[0, 99{,}999]$	1
Import Prices	float	$[0, 1]$	Defined by look-ahead
Export Prices	float	$[0, 1]$	Defined by look-ahead
Time of Day	int	$[0, 23]$	1
Predicted Community Production	float	$[0, 99{,}999]$	Defined by look-ahead
Predicted Community Consumption	float	$[0, 99{,}999]$	Defined by look-ahead

The action space, introduced in Table 2, is defined by a discrete control variable bounded between -1.0 and 1.0, with an interval of 0.1, which determines the type of action, discharge (negative values), idle (0.0), or charge (positive values), relative to the maximum discharge and charge limits. SoC values are kept between 0.2 and 1.0, with actions resulting in values outside of these being adjusted. When a BESS agent discharges, the amount (taking into account efficiencies) is added to the available energy pool and becomes visible to subsequent agents. Regarding import and export prices, a look-ahead window was employed, allowing agents to access information a fixed number of timesteps into the future. The specific look-ahead value is specified alongside other training parameters and is dependent on the environment's temporal resolution.

Table 2. BESS Action Space. Action values denote charge/discharge percentage relative to the maximum allowed charge/discharge.

Variable	Type	Bound	Step
Action	float	$[-1, 1]$	0.1

The costs considered for the BESS agents are determined based on the action taken, maximum allowed charge and discharge, and inherent monetary costs of each action. Furthermore, to ensure sufficient flexibility in the early hours of the following day, a penalty

is imposed when the SoC falls below a specified threshold at the end of the day. The formulations for the BESS agents' cost and penalty are presented in Equation (3) and Equation (4), respectively:

Equation (3) presents the formulation of the BESS agents' cost.

To ensure sufficient flexibility in the early hours of the following day, a penalty is imposed when the SoC falls below 50% at the end of the day, represented by γ.

$$\begin{aligned}\phi_{bess}(action) = \max(0, action) \cdot (\psi_{bess}\phi_{charge}) \\ + \max(0, -a) \cdot (\upsilon_{bess}\phi_{discharge})\end{aligned} \tag{3}$$

$$\rho_{bess}(action) = (\text{SoC}_{bess} < \gamma \cdot \text{Capacity}_{bess}) \cdot EoD \tag{4}$$

where $\phi_{bess}(action)$ is the cost of the *action* taken by the BESS agent, ψ_{bess} and υ_{bess} denote the maximum allowed charge and discharge limits, and ϕ_{charge} and $\phi_{discharge}$ represent the monetary costs of charging and discharging, respectively. Additionally, $\rho_{bess}(action)$ is the penalty applied to ensure reserve capacity, $textSoC_{bess}$ and Capacity_{bess} denote the current SoC and total capacity of the BESS, γ represents the end of day SoC threshold percentage, and *EoD* denotes the scalar penalty value applied should the threshold not be met.

This penalty encourages the agent to maintain reserve capacity, enhancing its ability to respond to demand during the specified time frame. A value of $EoD = 50$ (with γ set to 0.5) was used throughout this work. This value was chosen as it is high enough to avoid agents being too focused on the economic aspect, and prefer being stale as opposed to respect scheduled constraints. It is also low enough that it does not overwhelm the economic aspect and enables the intended charge and discharge behaviour, preventing the BESS agents from exclusively remaining above the end-of-day threshold at all times.

3.2.2. Electric Vehicles

The EVs can be considered an extension of the presented BESSs, with additional conditional variables that determine the grid connection and the power required for a trip. Consequently, the observation space also considers the grid connections of the EV over a specified horizon and in accordance with the trip power requirement. The observation space, along with the bounds used, is presented in Table 3. The current time was selected instead of the time of day to ensure that the information conveyed was on the same scale as the next departure time. The action space remains the same as the BESS, as presented in Table 2, and preserves the action correction strategy from the BESSs.

Table 3. EV Observation Space. Considers the added requirements of the EV to perform trips according to a schedule.

Variable	Type	Bounds	Shape
State of Charge	float	$[0, 1]$	1
Available Energy	float	$[-99{,}999, 99{,}999]$	1
Maximum Charge	float	$[0, 99{,}999]$	1
Maximum Discharge	float	$[0, 99{,}999]$	1
Grid Connections	int	$[0, 1]$	Defined by look-ahead
Next Departure Requirement	float	$[0, 1]$	1
Import Prices	float	$[0, 1]$	Defined by look-ahead
Export Prices	float	$[0, 1]$	Defined by look-ahead
Current Time	int	$[0, 9999]$	1
Predicted Community Consumption	float	$[0, 99{,}999]$	Defined by look-ahead
Predicted Community Production	float	$[0, 99{,}999]$	Defined by look-ahead

The reward structure employed closely resembles the BESS formulation, although it is not subject to a penalty based on their SoC at the end of the day. Instead, a penalty is

applied if an EV fails to meet the energy requirements for a scheduled trip in the current timestep. This guides agents toward vehicle availability while allowing flexibility during the remaining hours. The attribution of this penalty for EV agents is formulated in Equation (5) as follows:

$$\rho_{ev}(action) = (\text{SoC}_{ev} < \text{Requirement}_{trip}) \cdot TP \quad (5)$$

where $\rho_{ev}(action)$ is the resulting penalty applied to the EV agents, SoC_{ev} is the current vehicle SoC, $\text{Requirement}_{trip}$ denotes the energy required for a scheduled trip in the current timestep, and TP represents the scalar penalty assigned for failing to meet the requirement.

Following the same logic of the BESS agent formulation, the penalty value $TP = 50$ used was chosen to avoid EV behaviours that ignore mobility constraints and exclusively act as a storage unit for the community. However, it is not formulated as a hard constraint, as EVs should be able to not meet a mobility requirement if doing so would cause instability to the community.

3.2.3. Main Grid Connector

The objective of this component is to ensure that the total energy production matches the total energy consumption within the community, resulting in a net energy balance of zero. This component alludes to one of a community manager's roles, referred to as a connector for its role in interfacing with the external grid. To this end, a deterministic policy was implemented in which the agent exports energy when there is an excess and imports when there is a deficit. The cost associated with this energy balance is defined in Equation (6) as follows:

$$\phi_{ec} = \max(0, EB) \cdot \phi_{import} + \max(0, -EB) \cdot \phi_{export} \quad (6)$$

where ϕ_{ec} is the total energy cost for the connector, EB represents the available energy balance in the community, and phi_{import} and ϕ_{export} denote the monetary costs associated with importing and exporting energy, respectively.

Finally, a penalty is assessed whenever imports or exports exceed the maximum allowed values by multiplying the excess by a factor. This formulation is described in Equation (7):

$$\rho_{ec} = BP \cdot (\max(0, -EB - \psi_{ec}) + \max(0, EB - v_{ec})) \quad (7)$$

where ρ_{ec} is the balance penalty applied to the community, BP represents the scalar balance penalty factor, and ψ_{ec} and v_{ec} denote the maximum allowed import and export values, respectively. A value of $BP = 5000$ was set to avoid damaging the grid from excess energy (in the case of exports) and disconnections from insufficient energy (in the case of imports).

To promote training stability, a sparse reward structure was implemented, where agents received their reward only after all agents had acted. The reward formulation included the agent's cost, penalties, and a shared connector component to reflect community-wide performance. This approach attempts to avoid exploiting actions by a certain agent that, while optimal for said agent, are detrimental to the community. Sparse rewards can also be used to prevent premature convergence [40].

3.3. Priority Mechanisms

In addition to the proposed sequential MARL formulation, this work introduces two priority mechanisms: a fixed priority, where the order is set in advance, and an adaptive priority based on assets' contributions relative to the community's generation and production.

The fixed priority order—EVs, followed by BESSs, and finally the main grid connector—was adopted as a practical rule of thumb commonly observed in real-world energy resource management. Within each group, resources are prioritised based on maximum capacity, with

smaller resources given higher priority. This order was chosen to allow the EVs to respect scheduled trip requirements, while BESSs and the main grid connector provide flexibility to balance supply and demand as needed. Prioritisation based on lower capacity was chosen to incentivise the use of more resources at any given time.

The adaptive mechanism was adopted from [41], where resource priority is calculated by considering all contributions up to the current timestep for consumption and generation, relative to aggregated community values. By considering historical contributions, the mechanism adapts to varying consumption and generation patterns, providing a dynamic prioritisation process and enhancing efficiency.

A logarithmic transformation was applied to mitigate large numerical fluctuations resulting from high values (i.e., unusually high consumption or generation) and to help maintain stability and prevent outliers from overtaking the calculations. Similarly, an ϵ value was added to ensure positive, non-zero values for calculating priorities. The terms α and β were introduced to allow flexibility when setting community goals, where α represents the weight of the contributions in the form of generation, and β represents the weight assigned to the contributions in the form of consumption. In practice, the α and β terms enable community managers to adapt the mechanism to specific grid constraints and operational targets. For instance, in communities with high renewable density but restricted export capacity, an adoption of generation-first strategy ($\alpha > \beta$) inherently relates to the follow-up consumer agents that there is a possibility of using said energy, limiting curtailment and increasing self-consumption. Conversely, should there be a scenario where there are strict and specific peak-demand usage patterns, a consumer-first strategy ($\beta > \alpha$) can be used, providing the opportunity to consumer agents to meet their requirements early on. Finally, a balanced strategy ($\alpha = \beta$) serves a default scenario, where there is no favouring of generation or consumption.

Equation (8) showcases the described adaptive mechanism. As an intermediary to the main grid, the main grid connector does not follow this mechanism and is always considered the last acting agent to ensure the grid remains balanced, handling imports and exports as needed.

$$P_r = \alpha \times \frac{\log(generated_r + \epsilon)}{\log(generated_{ec} + \epsilon)} + \beta \times \frac{\log(consumed_r + \epsilon)}{\log(consumed_{ec} + \epsilon)} \quad (8)$$

4. Evaluation Methodology

This section presents necessary information on how the various approaches were trained and benchmarked. Section 4.1 details the dataset utilised, informing EV and BESS characteristics and how the profiles were generated. Training of the approaches is addressed in Section 4.2, which provides information on the libraries used and the training algorithms and parameters. Finally, Section 4.3 presents the experimental setup and how the benchmarks were conducted.

4.1. Dataset

The data utilised throughout this work are based on a community with hourly records spanning two years for five renewable energy sources and five loads, and can be found in [42]. Three BESSs were used, with Table 4 presenting the characteristics of each storage. Five EV profiles, found in Table 5, were generated using COTEV [43], with the maximum values for charging and discharging being dependent on the type of station (3.6 kWh or 7.2 kWh) used, depending on the calculated trip and the need to connect the vehicle mid-trip.

Table 4. Specifications of the BESS used.

Type	Capacity (kWh)	Charge Limit (kW)	Charge Efficiency (%)	Discharge Limit (kW)	Discharge Efficiency (%)
Li-ion	50	10	95	10	95
Li-ion	100	20	96	20	96
Li-ion	15	4	95	4	95

Table 5. Specifications of the EVs generated.

Type	Capacity (kWh)	Charge Efficiency (%)	Discharge Efficiency (%)	Energy Use (kWh/km)
PHEV	70.2	90	90	0.16
PHEV	13.5	90	90	0.18
PHEV	78.0	90	90	0.176
BEV	68.0	90	90	0.26
PHEV	70.2	90	90	0.16

4.2. Training

The methods used throughout this work were trained on one month of data (January 2019) and tested on the entire dataset. The priority mechanisms discussed were implemented and used during training. The defined look-ahead for agents was set to 12, effectively considering half a day's worth of look-ahead.

The Ray library (RLlib) was chosen as it provides a simple way to implement agent behaviour and already includes algorithms for training policies [44]. Furthermore, Ray permits only a subset of agents to have observations, enabling the proposed sequential approach. The Python version used was 3.11, and the Ray version used was 2.7.0. All the experiments in this work were conducted on an Apple M3 12-core chip running macOS Sequoia 15.2, and the code is available online (Code available on https://github.com/ECGomes/pyecom (accessed on 31 October 2026)).

The resulting policies were trained using the Proximal Policy Optimization (PPO) algorithm [45], with the training configuration presented in Table 6. PPO is an improvement over the Trust Region Policy Optimization (TRPO) algorithm, a method that attempts to navigate towards an optimal region of the solution space. PPO was designed to trim complex calculations, opting for simpler strategies to ensure policy updates are fast and feasible. Equation (9) presents PPO's update rule, where $r_t(\theta)$ denotes the probability ratio between the old and new policies (values higher than 1.0 represent a higher likelihood in the current policy), $\hat{A}_t$ represents the estimated advantage of choosing an action, ϵ denotes the clipping threshold [45]. The presence of both unclipped and clipped terms results in a flexible update that depends on the policy ratios and advantages.

The gamma values were set to 0.9 to strike a compromise between long-term and current rewards. Regarding the neural network architecture, two feed-forward layers were wrapped by an LSTM, and the model configuration is shown in Table 7. LSTMs were integrated to capture system dynamics over time, which is valuable in situations where actions at a given timestep influence future actions and improve overall learning stability. Seed values were fixed at 42.

Each agent was assigned a unique policy network, resulting in a small total number of parameters per policy. This approach was chosen as the learning complexity of a single agent is not altered by the variable total number of agents in the system.

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (9)$$

Table 6. Training configuration for RL approaches. Parameter names presented correspond to the RLlib nomenclature.

Parameter	Values
train_batch_size	256
lr	5×10^{-5}
gamma	0.99
use_gae	True
use_critic	True
use_kl_loss	True
clip_param	0.1
grad_clip	4
gamma (BESS)	0.9
gamma (EV)	0.9
gamma (Agg.)	0.9

Table 7. Trained model configuration for BESSs and EVs. vf_share_layers is set to False, per RLlib recommendations, whenever LSTM wrapping is enabled. Parameter names correspond to the names utilised in RLlib.

Parameter	Values
use_lstm	True
fcnet_hiddens	[32, 32]
lstm_cell_size	128
lstm_use_prev_action	True
lstm_use_prev_reward	True
vf_share_layers	False

4.3. Experimental Setup

This work presents an experiment to uncover the potential of a sequential MARL approach in ECs. To understand the impact of prioritisation in the considered system, two sequential approaches with different prioritisation mechanisms were compared to a simultaneous counterpart without any prioritisation. The sequential approaches used prioritisation based on (1) fixed agent order and (2) adaptive priority based on grid contributions. All strategies are compared using the resulting monthly costs. This design allows a controlled and fair assessment of the added value of the prioritisation mechanism itself.

Individual days were also inspected based on the adaptive approach's performance relative to the MARL baseline. The best- and worst-performing days are provided, along with an analysis of the decisions made and the resulting self-sufficiency and self-consumption values. The self-sufficiency was calculated following Equation (10), and the self-consumption was calculated following Equation (11). These values are reported for the best- and worst-case scenarios, to allow inspection of behavioural differences between sequential and simultaneous approaches under edge cases.

$$\text{Self-Sufficiency} = \frac{\text{Consumption} - \text{Imports}}{\text{Consumption}} \tag{10}$$

$$\text{Self-Consumption} = \frac{\text{Production} - \text{Exports}}{\text{Production}} \tag{11}$$

MILP and MPC were not considered as baselines due to the sequential approach relying on immediately available community information, under partial observability. While MILP offers optimality guarantees, it operates under full observability, which is typically not available in real-world deployments. As such, the comparisons drawn in this work are under the assumption of partial system observability, which is the case for every MARL approach explored.

5. Results and Discussion

This section presents the outcomes of the experiment conducted throughout this work, along with an interpretation of the results and a rationale for the observed behaviours. Section 5.1 presents the economic costs for each strategy, while Section 5.2 delves into the impacts of agent actions on the best- and worst-performing days in the dataset. Section 5.3 discusses findings related to the order, potential issues, and improvements.

5.1. Cost Comparison

This section presents the operational costs (€) of a sequential RL, with fixed and adaptive orders, and the multi-agent RL benchmark. The experimental results are presented in Table 8, with the training data indicated by an asterisk. The results presented are the sum of the costs at each timestep of the respective month, calculated from the agent's actions and their respective resource costs.

Table 8. Monthly and total costs (€) of compared approaches. The training month is highlighted with an asterisk (*) and is not included in the aggregate calculations. The differences relative to the baseline are also displayed as percentage values.

Year	Month	Adaptive	Sequential	Baseline	Lowest
2019	1 *	413.47 (−4.40%) *	423.73 (−2.03%) *	432.50 *	Adaptive *
	2	395.21 (−3.00%)	397.90 (−2.34%)	407.42	Adaptive
	3	394.89 (−3.59%)	401.39 (−2.01%)	409.61	Adaptive
	4	398.32 (−2.75%)	400.18 (−2.30%)	409.60	Adaptive
	5	393.15 (−2.30%)	398.58 (−0.95%)	402.42	Adaptive
	6	367.29 (−5.63%)	382.39 (−1.75%)	389.20	Adaptive
	7	384.67 (−2.75%)	384.35 (−2.83%)	395.55	Sequential
	8	418.97 (−2.61%)	420.44 (−2.27%)	430.19	Adaptive
	9	377.45 (−3.25%)	382.88 (−1.86%)	390.13	Adaptive
	10	405.42 (−3.50%)	410.94 (−2.19%)	420.14	Adaptive
	11	420.80 (−4.04%)	427.87 (−2.43%)	438.52	Adaptive
	12	357.40 (−6.44%)	366.83 (−3.97%)	381.99	Adaptive
2020	1	342.68 (−5.53%)	351.88 (−3.00%)	362.75	Adaptive
	2	408.85 (−4.87%)	415.92 (−3.22%)	429.78	Adaptive
	3	388.21 (−3.56%)	392.53 (−2.49%)	402.56	Adaptive
	4	306.96 (−7.03%)	311.19 (−5.75%)	330.17	Adaptive
	5	301.27 (−5.74%)	305.15 (−4.52%)	319.61	Adaptive
	6	319.08 (−4.45%)	319.75 (−4.25%)	333.95	Adaptive
	7	390.96 (−2.18%)	399.49 (−0.04%)	399.66	Adaptive
	8	384.08 (−3.61%)	387.60 (−2.73%)	398.47	Adaptive
	9	404.13 (−3.32%)	412.67 (−1.28%)	418.02	Adaptive
	10	378.45 (−4.66%)	388.55 (−2.11%)	396.94	Adaptive
	11	405.51 (−3.63%)	410.64 (−2.41%)	420.78	Adaptive
	12	409.58 (−4.46%)	413.46 (−3.55%)	428.68	Adaptive
$\bar{x}$		380.58 (−3.98%)	386.20 (−2.56%)	396.35	Adaptive
σ		33.89 (+5.12%)	34.10 (+5.77%)	32.24	Baseline
$\sum$		8753.33 (−3.98%)	8882.58 (−2.56%)	9116.14	Adaptive

Considering the costs in Table 8, the adaptive priority sequential approach consistently yields the lowest community costs, achieving the best results in twenty-three of the twenty-four months considered, with a marginal difference of 0.08% in the month when the fixed-order approach was best. This represents an improvement over the fixed sequential approach, which presents lower values than the baseline. These results highlight the potential of the proposed sequential approach when combined with priority mechanisms.

Depending on the month studied, the improvements achieved by the sequential strategies can yield savings of approximately 7% and 5.75% with adaptive and fixed priorities, respectively. Considering the entire dataset (including training data), the savings are 4.0% for the adaptive and 2.54% for the fixed-priority. Similarly, the savings obtained without considering the training data are 3.98% for the adaptive and 2.56% for the fixed-priority.

5.2. Daily Performance—Best and Worst

To provide more meaningful insights into the performance of the proposed sequential strategy, results for the best- and worst-performing days in the dataset are presented. The days were selected by dividing the adaptive approach costs by the baseline values and then calculating the maximum and minimum values. Data from the training set were not included in this calculation, and days with a minimum cost below the threshold of 5.0 were excluded to remove days that exhibit insufficient consumption and/or production. This process resulted in selecting the 5 of June 2020 as the best-performing day and the 9 of December 2020 as the worst-performing day. The results of the actions followed by agents for the best day are shown in Figure 4, and for the worst day in Figure 5.

Regarding the best day, the costs obtained for the approaches were 7.91 € for the adaptive priority, 7.88 € for the fixed priority, and 11.06 € for the baseline. While the most significant consumption spike occurred under the adaptive priority strategy, it was compensated for using BESS3 and EV3 earlier in the day. The fixed-priority approach leaned towards minimal resource use, achieving the lowest cost across the three approaches, charging EVs in small amounts throughout the day, with a small discharge at around six in the morning. Finally, the baseline also heavily relied on EVs, imposing greater charging requirements and leading to further import needs. The baseline approach also used BESS3, discharging during the morning. While resources were discharged across all approaches, almost no exports were recorded.

The worst-performing day incurred costs of 22.55 €, 19.88 €, and 19.55 € for the adaptive priority, fixed priority, and baseline, respectively. This day exhibits an atypical demand profile not present in the training data, with the adaptive priority strategy performing the worst and the baseline the best. The adaptive priority further exacerbated the issue, with agents charging both asset types during demand peaks. Fixed-priority and baseline approaches are similar in terms of consumption profiles, with the baseline showing slightly higher values, albeit supported by significant use of EV and BESS discharge actions. Although the adaptive priority approach incurred a higher cost on this day, the following week showed an average of 3.3% savings compared to the baseline, as shown in Table 9.

A deeper analysis of the impact of resource group costs is shown in Figures 6 and 7 for the best- and worst-performing days, respectively. On the best-performing day, the adaptive and fixed-priority approaches demonstrated marginally lower total costs by balancing BESS and EV. Conversely, on the worst-performing day, the baseline approach achieved the lowest costs by leveraging BESS and EV use. These results reveal that the baseline approach was more robust in edge cases and highlight the need to account for additional demand and production patterns when training sequential approaches.

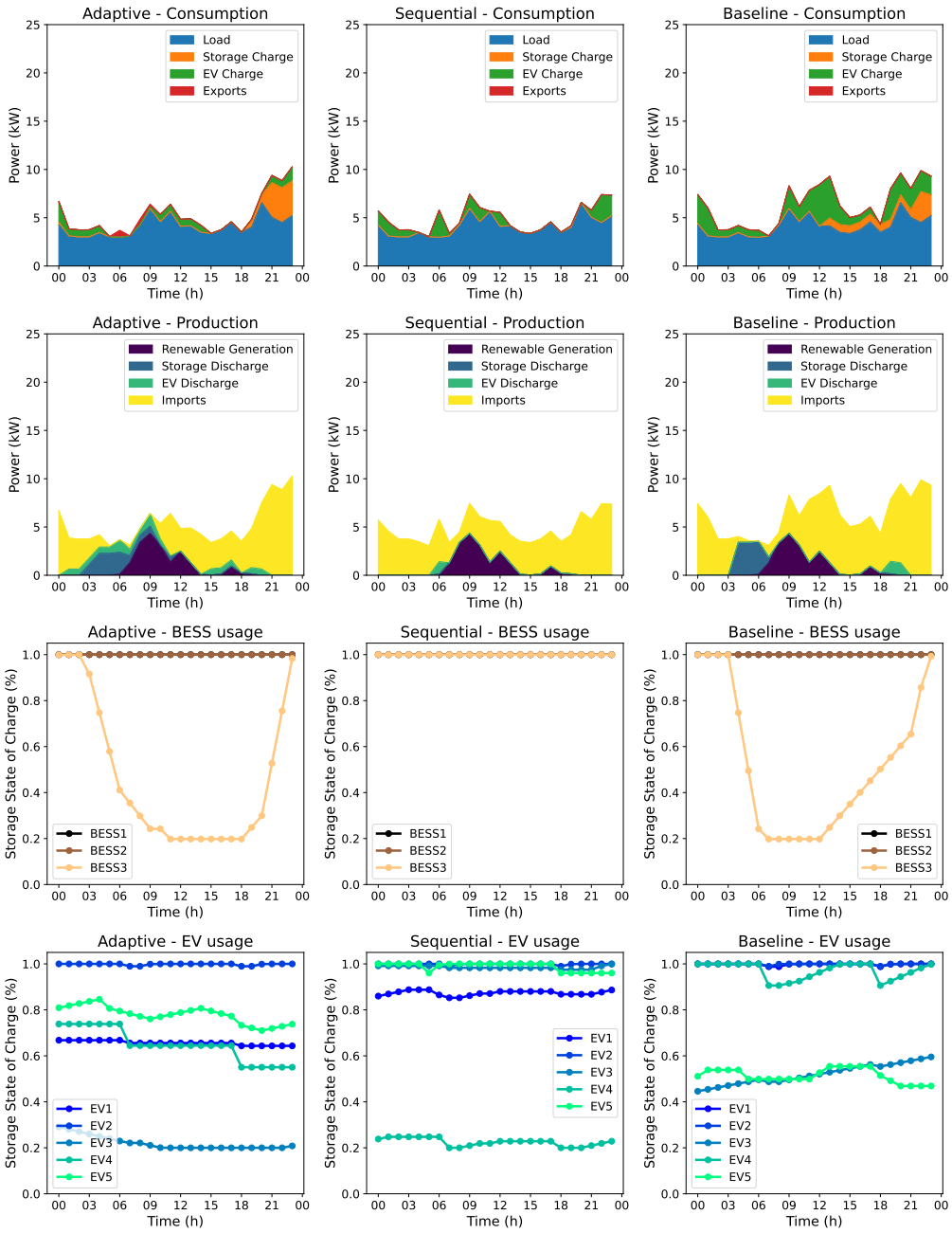


Figure 4. Best-performing day considering the adaptive approach in cost (€). Corresponds to the 5 of June 2020 in the dataset.

Table 9. Cost deviation of the adaptive method relative to the baseline following the worst-performing day. Values higher than 1 indicate a higher relative cost, while values lower than 1 indicate a lower relative cost.

Date	Deviation to the Baseline
2020-12-09	1.15
2020-12-10	0.95
2020-12-11	0.99
2020-12-12	0.88
2020-12-13	0.85
2020-12-14	1.08
2020-12-15	0.92
2020-12-16	0.91

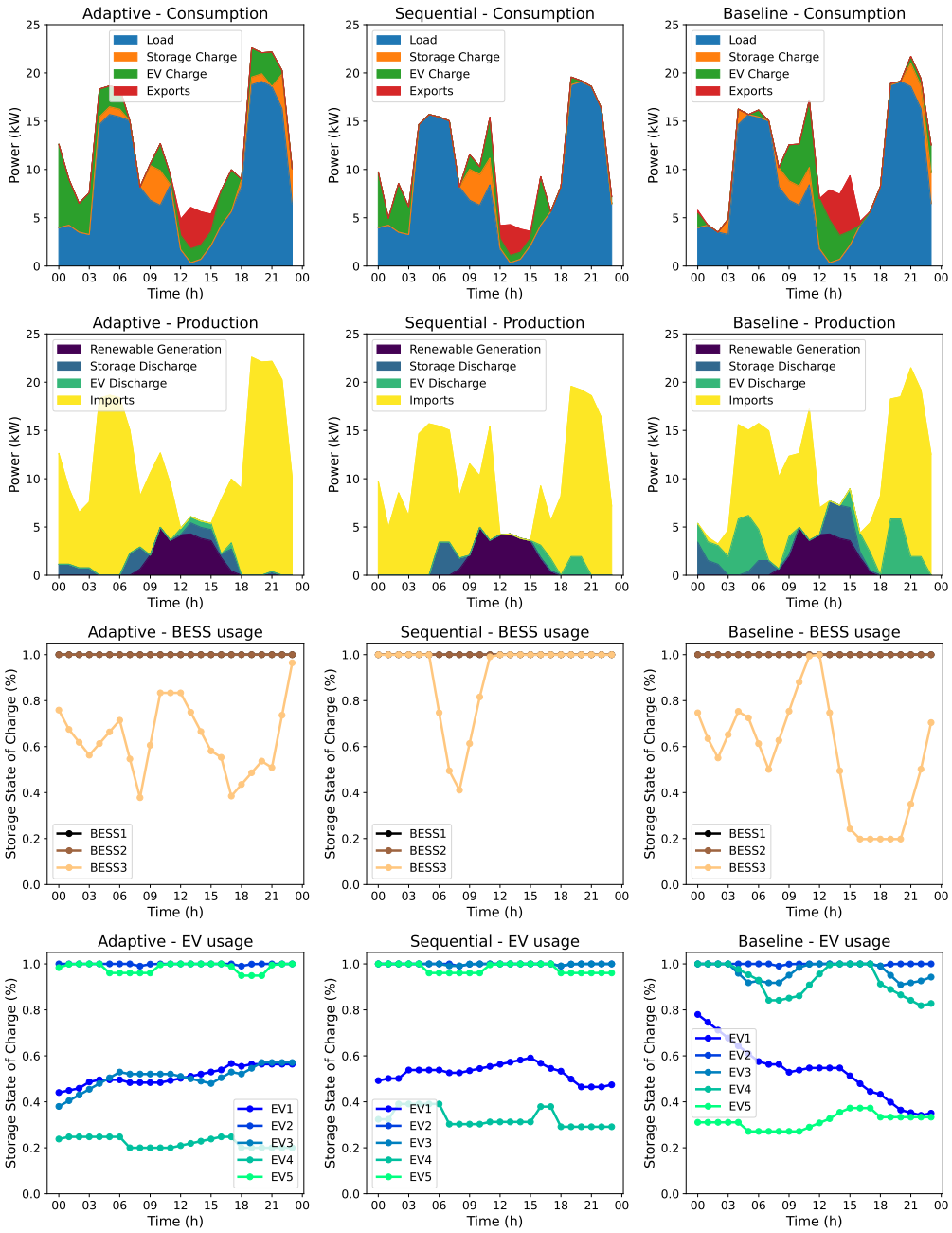


Figure 5. Worst-performing day considering the adaptive approach in cost (€). Corresponds to the 9 of December 2020 in the dataset.

Analysing self-sufficiency and self-consumption values (Table 10), adaptive priority exhibited marginally higher self-sufficiency (39%) compared to the fixed priority (16%) and baseline (27%). This difference can be attributed to the adaptive approach prioritising cost optimisation while considering resource impacts through contributions toward the community. Regarding self-consumption, all approaches performed similarly, with the adaptive priority achieving the lowest rate (99%) due to a small export amount.

Table 10. Self-sufficiency and self-consumption values for the best-performing day—5 June 2020.

	Adaptive	**Fixed-Priority**	**Baseline**
Self-Sufficiency	39%	16%	27%
Self-Consumption	99%	100%	100%

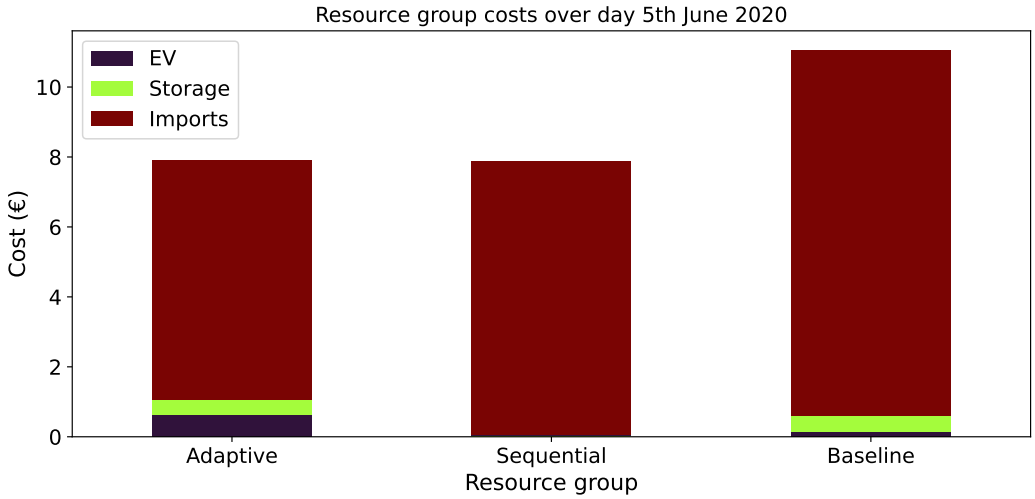


Figure 6. Resource group cost (€) for day 5th June 2020.

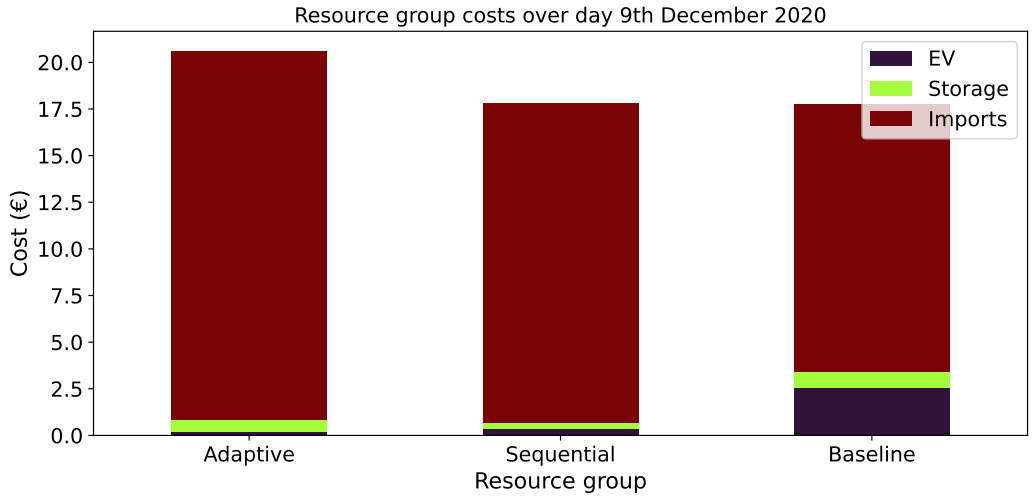


Figure 7. Resource group cost (€) for day 9th December 2020.

Concerning the worst-performing day, Table 11 provides information for the self-sufficiency and self-consumption values, indicating the baseline's superior performance regarding self-sufficiency (48%), with both adaptive and fixed-priority strategies demonstrating lower self-sufficiency (27%) due to higher import values. Self-consumption was relatively high across adaptive, fixed-priority, and baseline methods (92% to 93%), with the baseline slightly outperforming the remaining approaches.

Table 11. Self-sufficiency and self-consumption values for the worst-performing day—9 December 2020.

	Adaptive	Fixed-Priority	Baseline
Self-Sufficiency	27%	27%	48%
Self-Consumption	92%	92%	93%

Finally, the observed reduced performance during the worst-performing day further suggests that adaptive methods were not fully prepared to tackle daily patterns outside the distribution in which they were trained. The use of deterministic EV scheduling should also be taken into account, as EV scheduling carries a significant impact on the allocation of community resources [46–48]. The inclusion of other seasonal profiles, stochastic EV behaviour,

and expanded training data would further benefit the proposed methodology by allowing exposure to more atypical patterns. A hybrid approach to detect out-of-distribution patterns, such as dynamically altering the prioritisation mechanism, could also be considered as a possible research direction to build more robust solutions.

5.3. Effects of the Ordering Mechanisms

Considering the adaptive priority mechanism, a total of 117 unique priority sequences emerged throughout the two-year dataset. Figure 8 illustrates the shifts in priorities, assigning integer identifiers to each distinct sequence. Although priorities were reset at the beginning of each month, there was a consistent convergence toward a stable set of orders within a few days, typically maintaining the order until the next reset. The three most frequently occurring sequences were:

- [EV5, BESS3, EV3, EV1, EV4, EV2, BESS1, BESS2] (37%);
- [EV5, EV3, BESS3, EV1, EV4, EV2, BESS1, BESS2] (29%);
- [EV5, BESS3, EV3, EV4, EV1, EV2, BESS2, BESS1] (7%).

In contrast, the fixed-priority approach utilized the sequence [EV2, EV4, EV1, EV5, EV3, BESS3, BESS1, BESS2].

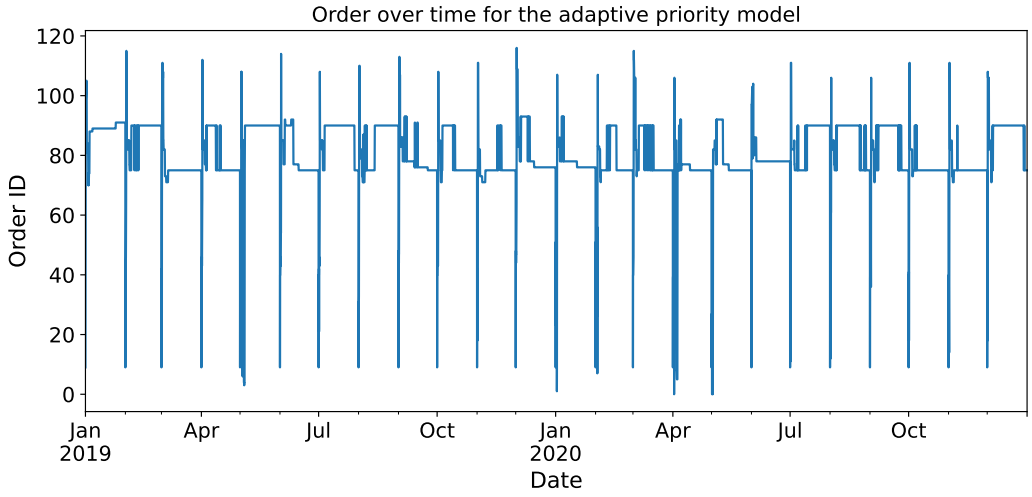


Figure 8. Order IDs utilised throughout the dataset. Every unique order was assigned an integer to represent the evolution over time.

The frequent appearance of EV5 and BESS3 among the top-priority agents highlights the role of these assets in managing community resources, likely due to their substantial contributions or capacity characteristics. Notably, EV5, despite having the second-largest trip power requirements, often received the highest priority, whereas EV4, having the highest power requirement, typically occupied mid-level positions. Similarly, EV3, with the largest capacity, regularly appeared near the top of the priority order. This observation differs from the initial priority principle, in which lower-capacity resources were given precedence. However, the consistent positioning of BESS3 as the highest-priority BESS aligns with the initial prioritisation approach.

Given these observations, recalculating priorities more frequently or using mechanisms such as moving averages or sliding windows could improve the adaptive mechanism's responsiveness to community dynamics and increase efficiency.

6. Conclusions

In this work, a novel sequential MARL approach was proposed and studied for resource management within ECs. While traditional MARL strategies allow simultaneous decision-making and can lead to resource conflicts and suboptimal optimisation, the proposed approach utilises sequential decision-making, in which each agent has access to the most up-to-date environmental information.

The proposed method was evaluated against a MARL approach using operational cost metrics across two years of data. Results demonstrated that the sequential approach consistently outperformed the baseline, showing an average cost saving of 2.57%. Additionally, a prioritisation mechanism was proposed to further demonstrate the applicability of the method. With an adaptive priority mechanism based on contribution, costs were further reduced, resulting in an average cost reduction of 4%. An analysis of the best- and worst-performing days highlights the potential of the proposed method while revealing limitations that can be addressed in future research. Notably, in scenarios where the sequential methods outperform the benchmark, the margin of improvement is greater compared to when they underperform.

Despite promising results, there are also some limitations to address in future iterations of this work. The training data, limited to one month, may restrict agent adaptability to unforeseen scenarios, as seen on the worst-performing day. Extending the training period, incorporating diverse demand profiles and integrating stochastic EV profile behaviour could enhance robustness and capture unobserved system dynamics. Additionally, the MARL framework assumes fixed agent characteristics and efficiencies, which limits its applicability under uncertainty or long-term degradation. Future work could explore adaptive models that account for these factors.

Furthermore, the proposed setup adopts a simplified representation of an EC, focusing on a limited number of homogeneous assets and fixed pricing conditions. This choice follows common practice in the literature, where reduced-scale or simplified configurations are used to evaluate control strategies with clarity and ease [8,49]. However, while such assumptions enable the isolation of the effects of sequential decision-making, future work should extend the analysis to a larger number of participants, agent objectives, dynamic pricing, and uncertain forecasts to assess scalability and robustness under more realistic conditions. Environment assumptions, such as using clipping mechanisms for battery behaviour and continuous action spaces, should also be considered as possible improvements to training stability.

Additionally, the scope of this work does not include an in-depth exploration of novel priority mechanisms. Instead, two well-established strategies, fixed and adaptive priority attribution, are used to evaluate how such mechanisms interact with sequential MARL. This choice allows isolation of the impact of decision ordering itself, rather than introducing new heuristics, providing a clear foundation for future research on more complex or context-aware priority designs.

Moreover, while the Results Section discusses agent orderings from a technical standpoint, the study does not delve into fairness or user-centric performance metrics. Since these are key concerns for real-world EC development, future work should also explore how priority mechanisms can be adapted to support these goals. Moreover, while this work focuses on standard MARL settings, future research could investigate how advanced paradigms, such as mean-field or federated Reinforcement Learning, might benefit from or integrate priority-based sequential mechanisms.

Finally, the adaptive priority mechanism relies on cumulative contributions without special consideration for recent performance, potentially leading to suboptimal prioritisation. Introducing a sliding window or periodic resets could improve responsiveness in dynamic

scenarios. Alternative priority-sorting mechanisms, such as those incorporating social factors, also present avenues for further research.

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

Funding: This work was possible with the support of Fundação para a Ciência e a Tecnologia, I.P. (FCT) under scholarship 2021.07754.BD (https://doi.org/10.54499/2021.07754.BD), and projects UID/50021/2025 (https://doi.org/10.54499/UID/50021/2025), UID/PRR/50021/2025 (https://doi.org/10.54499/UID/PRR/50021/2025), 10.54499/UIDP/50009/2025 (https://doi.org/10.54499/UID/50009/2025), and 10.54499/UIDB/50009/2025 (https://doi.org/10.54499/UID/50009/2025). This work also received funding from project n.º 56—"ATE", financed by European Funds, namely the "Recovery and Resilience Plan—Component 5: Agendas Mobilizadoras para a Inovação Empresarial", included in the NextGenerationEU funding program. Additionally, this work has been developed under the activities of the EV4EU and U2Demo projects, funded by the European Union's Horizon Europe research and innovation programme under grant agreement no. 101056765 and no. 101160684, respectively. Views and opinions expressed in this document are those of the authors only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: During the preparation of this work the authors used Grammarly (version 1.205.1.0) in order to review the text and improve clarity and coherence. After using this tool/service, the authors reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

BESS	Battery Energy Storage System.
CEC	Citizen Energy Community
EC	Energy Community
EV	Electric Vehicle
HRL	Hierarchical Reinforcement Learning
MARL	Multi-Agent Reinforcement Learning
MILP	Mixed-Integer Linear Programming
ML	Machine Learning
MPC	Model Predictive Control
P2P	Peer-to-Peer
PPO	Proximal Policy Optimization
REC	Renewable Energy Community
RES	Renewable Energy Source
RL	Reinforcement Learning
SoC	State of Charge
TRPO	Trust Region Policy Optimization

References

1. Hafez, F.S.; Sa'di, B.; Safa-Gamal, M.; Taufiq-Yap, Y.H.; Alrifaey, M.; Seyedmahmoudian, M.; Stojcevski, A.; Horan, B.; Mekhilef, S. Energy Efficiency in Sustainable Buildings: A Systematic Review with Taxonomy, Challenges, Motivations, Methodological Aspects, Recommendations, and Pathways for Future Research. *Energy Strategy Rev.* **2023**, *45*, 101013. [CrossRef]
2. European Commission. *2050 Long-Term Strategy*; European Commission: Brussels, Belgium, 2020.
3. Nijsse, F.J.M.M.; Mercure, J.F.; Ameli, N.; Larosa, F.; Kothari, S.; Rickman, J.; Vercoulen, P.; Pollitt, H. The momentum of the solar energy transition. *Nat. Commun.* **2023**, *14*, 6542. [CrossRef] [PubMed]
4. Moroni, S.; Alberti, V.; Antoniucci, V.; Bisello, A. Energy communities in the transition to a low-carbon future: A taxonomical approach and some policy dilemmas. *J. Environ. Manag.* **2019**, *236*, 45–53. [CrossRef] [PubMed]
5. Uihlein, A.; Caramizaru, A. *Energy Communities: An Overview of Energy and Social Innovation*; Publications Office of the European Union: Luxembourg, 2020.
6. Charbonnier, F.; Morstyn, T.; McCulloch, M.D. Coordination of resources at the edge of the electricity grid: Systematic review and taxonomy. *Appl. Energy* **2022**, *318*, 119188. [CrossRef]
7. Ghasemnejad, H.; Rashidinejad, M.; Abdollahi, A.; Dorahaki, S. Energy management in citizen energy communities: A flexibility-constrained robust optimization approach considering prosumers comfort. *Appl. Energy* **2024**, *356*, 122456. [CrossRef]
8. Guiducci, L.; Palma, G.; Stentati, M.; Rizzo, A.; Paoletti, S. A Reinforcement Learning approach to the management of Renewable Energy Communities. In Proceedings of the 2023 12th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 6–10 June 2023; pp. 1–8. [CrossRef]
9. Ding, Z.; Liu, Z.; Fang, Z.; Su, K.; Zhu, L.; Lu, Z. Multi-Agent Coordination via Multi-Level Communication. *arXiv* **2024**, arXiv:2209.12713. [CrossRef]
10. Volpato, G.; Carraro, G.; Cont, M.; Danieli, P.; Rech, S.; Lazzaretto, A. General guidelines for the optimal economic aggregation of prosumers in energy communities. *Energy* **2022**, *258*, 124800. [CrossRef]
11. Srithapon, C.; Månsson, D. Predictive control and coordination for energy community flexibility with electric vehicles, heat pumps and thermal energy storage. *Appl. Energy* **2023**, *347*, 121500. [CrossRef]
12. Lee, W.P.; Han, D.; Won, D. Grid-Oriented Coordination Strategy of Prosumers Using Game-theoretic Peer-to-Peer Trading Framework in Energy Community. *Appl. Energy* **2022**, *326*, 119980. [CrossRef]
13. Ding, L.; Ke, S.; Yang, J.; Shi, X.; Fan, P.; Crisostomi, E. A heuristic-fuzzy improved virtual synchronous generator control strategy for charging station frequency regulation. *J. Energy Storage* **2024**, *99*, 113411. [CrossRef]
14. Yan, L.; Chen, X.; Chen, Y.; Wen, J. A Hierarchical Deep Reinforcement Learning-Based Community Energy Trading Scheme for a Neighborhood of Smart Households. *IEEE Trans. Smart Grid* **2022**, *13*, 4747–4758. [CrossRef]
15. Chu, Y.; Wei, Z.; Fang, X.; Chen, S.; Zhou, Y. A Multiagent Federated Reinforcement Learning Approach for Plug-In Electric Vehicle Fleet Charging Coordination in a Residential Community. *IEEE Access* **2022**, *10*, 98535–98548. [CrossRef]
16. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; Adaptive Computation and Machine Learning Series; The MIT Press: Cambridge, MA, USA, 2018.
17. Zou, L.; Munir, M.S.; Tun, Y.K.; Kang, S.; Hong, C.S. Intelligent EV Charging for Urban Prosumer Communities: An Auction and Multi-Agent Deep Reinforcement Learning Approach. *IEEE Trans. Netw. Serv. Manag.* **2022**, *19*, 4384–4407. [CrossRef]
18. Talihati, B.; Fu, S.; Zhang, B.; Zhao, Y.; Wang, Y.; Sun, Y. Community shared ES-PV system for managing electric vehicle loads via multi-agent reinforcement learning. *Appl. Energy* **2025**, *380*, 125039. [CrossRef]
19. Liu, X.; Fan, P.; Yang, J.; Tahmeed, U. A flexible frequency control strategy for airport-microgrid through vehicle-to-grid and aviation-to-grid. *J. Energy Storage* **2025**, *125*, 116980. [CrossRef]
20. Shi, Y.; Dong, H.; He, C.R.; Chen, Y.; Song, Z. Mixed Vehicle Platoon Forming: A Multiagent Reinforcement Learning Approach. *IEEE Internet Things J.* **2022**, *12*, 16886–16898. [CrossRef]
21. 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. Top. Comput. Intell.* **2022**, *6*, 568–579. [CrossRef]
22. Zhou, S.; Hu, Z.; Gu, W.; Jiang, M.; Zhang, X.P. Artificial intelligence based smart energy community management: A reinforcement learning approach. *CSEE J. Power Energy Syst.* **2019**, *5*, 1–10. [CrossRef]
23. Ye, Y.; Tang, Y.; Wang, H.; Zhang, X.P.; Strbac, G. A Scalable Privacy-Preserving Multi-Agent Deep Reinforcement Learning Approach for Large-Scale Peer-to-Peer Transactive Energy Trading. *IEEE Trans. Smart Grid* **2021**, *12*, 5185–5200. [CrossRef]
24. Ye, Z.; Qiu, D.; Li, S.; Fan, Z.; Strbac, G. Federated Reinforcement Learning for decentralized peer-to-peer energy trading. *Energy AI* **2025**, *20*, 100500. [CrossRef]
25. Hashemnezhad, M.; Delkhosh, H.; Parsa Moghaddam, M. Aggregator pricing strategy for community energy management based on multi-agent reinforcement learning considering customer loss or gain. *Sustain. Energy Grids Netw.* **2025**, *41*, 101607. [CrossRef]

26. Singh, A.; Panigrahi, B.K. Prosumer Cost Efficiency and Ensuring Grid Stability Through a Hierarchical PPO Framework in Decentralized Community Energy Management. *IEEE Access* **2025**, *13*, 38368–38386. [CrossRef]
27. Singh, A.R.; Seshu Kumar, R.; Bajaj, M.; Hemanth Kumar, B.; Blazek, V.; Prokop, L. A blockchain-enabled multi-agent deep reinforcement learning framework for real-time demand response in renewable energy grids. *Energy Strategy Rev.* **2025**, *62*, 101905. [CrossRef]
28. Jendoubi, I.; Bouffard, F. Multi-agent hierarchical reinforcement learning for energy management. *Appl. Energy* **2023**, *332*, 120500. [CrossRef]
29. Tsaousoglou, G.; Giraldo, J.S.; Paterakis, N.G. Market Mechanisms for Local Electricity Markets: A review of models, solution concepts and algorithmic techniques. *Renew. Sustain. Energy Rev.* **2022**, *156*, 111890. [CrossRef]
30. Gokcek, T.; Sengor, I.; Erdinc, O. A Novel Multi-Hierarchical Bidding Strategy for Peer-to-Peer Energy Trading Among Communities. *IEEE Access* **2022**, *10*, 23798–23807. [CrossRef]
31. Wu, Y.; Zhao, T.; Yan, H.; Liu, M.; Liu, N. Hierarchical Hybrid Multi-Agent Deep Reinforcement Learning for Peer-to-Peer Energy Trading Among Multiple Heterogeneous Microgrids. *IEEE Trans. Smart Grid* **2023**, *14*, 4649–4665. [CrossRef]
32. Zang, Y.; He, J.; Li, K.; Fu, H.; Fu, Q.; Xing, J. Sequential Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23*; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 2023; pp. 485–493.
33. Wan, X.; Yang, C.; Yang, C.; Song, J.; Sun, M. SrSv: Integrating Sequential Rollouts with Sequential Value Estimation for Multi-agent Reinforcement Learning. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 23333–23342. [CrossRef]
34. Wang, X.; Tian, Z.; Wan, Z.; Wen, Y.; Wang, J.; Zhang, W. Order Matters: Agent-by-agent Policy Optimization. *arXiv* **2022**, arXiv:2302.06205.
35. Yang, T.; Shi, X.; Yang, Y.; Xu, C.; Liu, H.; Zeng, Q. Priority Sequential Inference: Improving Safety for Efficient Autonomous Highway Driving Using MARL. *IEEE Sens. J.* **2025**, *25*, 10390–10401. [CrossRef]
36. Zhang, B.; Li, L.; Xu, Z.; Li, D.; Fan, G. Inducing stackelberg equilibrium through spatio-temporal sequential decision-making in multi-agent reinforcement learning. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23, Macao, China, 19–25 August 2023; pp. 353–361. [CrossRef]
37. Bilakanti, N.; Gurung, N.; Chen, H.; Kothandaraman, S.R. Priority-based Management Algorithm in Distributed Energy Resource Management Systems. In Proceedings of the 2021 IEEE Green Technologies Conference (GreenTech), Denver, CO, USA, 7–9 April 2021; pp. 351–356. [CrossRef]
38. Ruan, L.; Yan, Y.; Guo, S.; Wen, F.; Qiu, X. Priority-Based Residential Energy Management With Collaborative Edge and Cloud Computing. *IEEE Trans. Ind. Inform.* **2020**, *16*, 1848–1857. [CrossRef]
39. Mussadiq, U.; Ahmed, S.; Gul, N.; Kim, J.; Kim, S.M. Priority-Based Energy Sharing and Management Among Prosumers in Smart Grids. *IEEE Access* **2022**, *10*, 12179–12190. [CrossRef]
40. Vecerik, M.; Hester, T.; Scholz, J.; Wang, F.; Pietquin, O.; Piot, B.; Heess, N.; Rothörl, T.; Lampe, T.; Riedmiller, M. Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewards. *arXiv* **2018**, arXiv:1707.08817. [CrossRef]
41. Jadhav, A.M.; Patne, N.R. Priority-Based Energy Scheduling in a Smart Distributed Network With Multiple Microgrids. *IEEE Trans. Ind. Inform.* **2017**, *13*, 3134–3143. [CrossRef]
42. Pereira, L.; Cavaleiro, J.; Barros, L. Economic Assessment of Solar-Powered Residential Battery Energy Storage Systems: The Case of Madeira Island, Portugal. *Appl. Sci.* **2020**, *10*, 7366. [CrossRef]
43. Guzman, C.P.; Gomes, E.; Pereira, L.; Morais, H. A Computational Implementation for Creating Electric Vehicles Profiles. In Proceedings of the 2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), Grenoble, France, 23–26 October 2023; pp. 1–5. [CrossRef]
44. Liang, E.; Liaw, R.; Nishihara, R.; Moritz, P.; Fox, R.; Goldberg, K.; Gonzalez, J.; Jordan, M.; Stoica, I. RLlib: Abstractions for Distributed Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2018; pp. 3053–3062.
45. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
46. Chen, X.; Zhang, T.; Ye, W.; Wang, Z.; Iu, H.H.C. Blockchain-Based Electric Vehicle Incentive System for Renewable Energy Consumption. *IEEE Trans. Circuits Syst. II Express Briefs* **2020**, *68*, 396–400. [CrossRef]
47. Zhang, T.; Chen, X.; Wu, B.; Dedeoglu, M.; Zhang, J.; Trajkovic, L. Stochastic Modeling and Analysis of Public Electric Vehicle Fleet Charging Station Operations. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 9252–9265. [CrossRef]

48. Chen, X.; Wang, H.; Wu, F.; Wu, Y.; González, M.C.; Zhang, J. Multimicrogrid Load Balancing Through EV Charging Networks. *IEEE Internet Things J*. **2021**, *9*, 5019–5026. [CrossRef]
49. Faia, R.; Soares, J.; Pinto, T.; Lezama, F.; Vale, Z.; Corchado, J.M. Optimal Model for Local Energy Community Scheduling Considering Peer to Peer Electricity Transactions. *IEEE Access* **2021**, *9*, 12420–12430. [CrossRef]

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.