energies

MDPI

Article

Energy Management Simulation with Multi-Agent Reinforcement Learning: An Approach to Achieve Reliability and Resilience

Kapil Deshpande *,†, Philipp Möhl, Alexander Hämmerle, Georg Weichhart, Helmut Zörrer and Andreas Pichler

Profactor GmbH, Robotics and Automation Systems Department, 4407 Steyr, Austria
* Correspondence: kapil.deshpande@profactor.at
† Current address: Profactor GmbH, Im Stadtgut D1 Steyr-Gleink, Upper Austria, 4407 Steyr, Austria.

Abstract: The share of energy produced by small-scale renewable energy sources, including photo-voltaic panels and wind turbines, will significantly increase in the near future. These systems will be integrated in microgrids to strengthen the independence of energy consumers. This work deals with energy management in microgrids, taking into account the volatile nature of renewable energy sources. In the developed approach, *Multi-Agent Reinforcement Learning* is applied, where agents represent microgrid components. The individual agents are trained to make good decisions with respect to adapting to the energy load in the grid. Training of agents leverages the historic energy profile data for energy consumption and renewable energy production. The implemented energy management simulation shows good performance and balances the energy flows. The quantitative performance evaluation includes comparisons with the exact solutions from a linear program. The computational results demonstrate good generalisation capabilities of the trained agents and the impact of these capabilities on the reliability and resilience of energy management in microgrids.

Keywords: energy management; multi-agent reinforcement learning; renewable energy systems; microgrid

Citation: Deshpande, K.; Möhl, P.; Hämmerle, A.; Weichhart, G.; Zörrer, H.; Pichler, A. Energy Management Simulation with Multi-Agent Reinforcement Learning: An Approach to Achieve Reliability and Resilience. *Energies* **2022**, *15*, 7381. https://doi.org/10.3390/en15197381

Academic Editors: Pedro Faria and Ramiro Barbosa

Received: 29 August 2022
Accepted: 23 September 2022
Published: 8 October 2022

1. Introduction

Curtailing man-made contributions to climate change is one of the biggest challenges of the 21st century. To this end it is pivotal to leverage renewable energy sources.

Renewable energy sources increase the stochastic effects on the supply side. Additionally, the demand side faces changes. This includes new types of consumers, such as electrical vehicles, and agile, on-demand production systems, such as additive manufacturing systems. In addition to this, new prosumers, such as pumped storage power plants, supercapacitors [1], and photo-voltaic (PV) attached batteries [2], are on the rise.

In general, fluctuations on the demand and consumer side will put energy networks under much more stress. The introduced dynamics is a threat, with respect to blackouts, increasing the need for preparedness for such events. Energy intensive industries, such as steel production, have been using their own small power plants to fulfil their energy needs in-short, becoming independent from the main grid [3]. Following this approach, other industries also wish to become less dependent. For example, a fish farm in Mali has benefited from using a PV microgrid for maintaining continuous power supply to their water treatment plant as "a 30-min power outage would mean the death of the fish in the pools, resulting in huge financial losses" [4]. The Austrian company Fronius [4] shows that the "solar system is providing 98% of firm requirements while 2% is generated with backup diesel generator". This does not only improve the resiliency of the farm but is also reducing its carbon footprint.

For rural electrification, having Renewable Energy (RE) grids involving PV panels is a good way of generating green energy, but "these mini-grids require investments in a rather complex power generation and distribution infrastructure" [5]. Additionally, the production is heavily dependent on the location. In a European region, the weather changes a lot (in contrast to, e.g., Africa) and roof construction styles and location of the industry brings more complexity to the overall system.

The above described dynamics put extreme stress on grids. Attention needs to be placed on energy management to avoid a breakdown of the overall system. In this applied research, our goal was to increase the resilience and reliability of microgrids by means of effective energy management. The vision is to develop a method that could be deployed decentralised, controlling the reactions of systems in the grid in order to maintain the balance of energy produced and consumed.

The first step towards this vision is to analyse data-driven methods that can be trained and later deployed in a decentralised manner, enabling systems to independently react to disturbances in the grid. The initially developed simulation environment allowed us to research such methods without real-time constraints.

We adapted a Multi-Agent Reinforcement Learning approach and trained the agents centrally. However, it is possible to deploy the agents decentralised. We used real world consumer and producer energy data for the development. Industrial users have provided input with respect to energy management systems and future microgrids. To improve the generalisation capabilities of the trained multi-agent reinforcement learning model, a generalised method was developed. A method to calculate a resilience and reliability score for a microgrid was developed. Quantitative results show that the generalised training method improves the resilience and reliability of the trained EM simulation, compared to simulation models that have been trained with specific energy data.

The rest of the paper is structured as follows. Section 2 gives an overview of the related work and discusses the key contributions of this article, Section 3 describes an approach to solve the real world problem by defining a problem statement, understanding the data and the simulation, and also defining an evaluation criteria for both. Finally, Sections 4 and 5 provide the results and conclusions of our research.

2. Related Work

2.1. Micro Grids

In a *smart grid*, as an extended concept of an intelligent power grid, individual components (storage, producers, consumers) of the energy network can communicate and coordinate with each other [6]. This forms the technical basis that allows one to achieve a producer/consumer balance between the components. This is important to limit the effects of extreme weather and other events with low probability and high impact on power systems, which have become increasingly evident worldwide over the last decades [7].

Microgrids represent a subgroup of smart grids that can also be operated autonomously by including locally generated (renewable) energy to increase resilience [8] According to IEEE standard 2030.7, a microgrid is *"a group of interconnected loads and Distributed Energy Resources (DER) with clearly defined electrical boundaries that acts as a single controllable entity with respect to the grid. It can connect and disconnect from the grid to enable operation in both grid-connected or island modes"* [9] (p. 13).

Some other research on microgrids is focusing on multi-energy microgrids [10] and megagrids [11].

Due to the current challenge of reducing CO_2 and thus the increasing need to use renewable energies (solar, wind, etc.), as well as the progress in the field of electrical energy storage, research on multigrid/multi-energy systems is becoming more and more important. Abu Elzait et al. [12] observed a drop in the prices of PV systems and showed in their studies that renewable-based microgrids are more economical than microgrids that utilize only conventional energy systems. Energy storage is an important aspect in

efficient microgrids. Machine learning can support the energy management of microgrids by leveraging different technologies and strategies [13].

2.2. Reliability and Resilience

Weather induced events, in general, lead to High Probability, Low Impact (HPLI) events [14]. In recent years, not only climate induced energy uncertainty (CinU) have led to Low Probability, High Impact (LPHI) events. Important characteristics of energy systems are their reliability and resilience [15].

There is no common definition of reliability and resilience in general, and with respect to microgrids in particular. In the following, we review work that addresses one of the two system properties—or even both.

With respect to microgrids, the difference between resilience and reliability can be determined by survivability, where survivability is defined as the ability of the system to maintain its supply even at a degraded level to consumers during a disturbance [16]. Reliability deals with preparedness against HPLI events, whereas resilience deals with preparedness against LPHI events. Resilience describes the ability to survive and quickly recover from extreme and unexpected disruptions [17].

For a system to be resilient, it either has to be robust, so that disturbances have less to no impact, or it can adapt quickly to get back to an output as it was before the event [18].

Resilient microgrids can be built by supporting an island mode. Here microgrids are trying to continue to operate independently without the main grid. For example, a microgrid in Maryland was able to supply its local loads by islanding from the grid during Superstorm Sandy in the USA [19].

Reliability can be defined as the ability of the power system to deliver electricity in the quantity and with the quality demanded by the users [20]. In short, the reliability of energy systems means that the lights are always "on" in a consistent manner.

Cuadra et al. [21] summarized the differences of concept resilience vs. reliability in power grids as: Resilience is related to LPHI events. It is a dynamic concept. Reliability is related to HPLI events. It is a static concept.

Amani et al. [22] depicted a literature review of different methods of measuring reliability and resilience of power grids and compared the performance of different metrics when applied to scenarios in benchmarks and real power grids.

Wang et al. [23] discussed an approach of how microgrids can help increase resilience. With respect to the operational strategies used to improve resilience, they classified the research areas of interest as: network reconfiguration, maingrid islanding, feasible islanding, demand side response, and vulnerability analysis.

Panteli et al. [7] described in their work the *multi-phase resilience trapezoid* phases (disturbance progress, post-disturbance degraded, and restorative) and metrics during a disaster event and its restoration.

Mujjuni et al. [24] presented a framework that links resilience to development states within the Electricity Supply Industry (ESI). They proposed 303 resilience indicators linked to 13 development goals, measured against 6 capacities and 11 qualities.

For prevention from extreme weather conditions, e.g., Panteli et al. [25] evaluated the relationship between windstorms and the failure probability of transmission components by using a mix of infrastructure and operational indices. Dehghani et al. [26] optimized the preventive replacement of poles in a large-scale power distribution system. Jufri et al. [27] studied in detail different grid resilience indices and classified grid resilience enhancement strategies into the groups *physical hardiness* and *operational capability*. Resilience enhancement strategies were categorized by Huang et al. [28] into resilience planning, resilience response, and resilience restoration, and they argue that the ultimate goal of system resilience enhancements leads to smart grids. This also leads to a more specific field of application: the recovery from cyber-attacks in smart grids [29].

Different measures can be applied to minimize the impact of the disturbance and decrease the time to restore the state [30].

2.3. Reinforcement Learning and Energy Management System

Active Energy Management (EM) has to be applied to increase resilience.

In recent years, the application of Reinforcement Learning (RL) for solving EM problems has significantly increased [8,31]. A few select research problems are found in the list below [8,32]:

- Energy management (energy cost, load peaks, electricity balance, etc.);
- Load and demand forecasting;
- Demand response (total profits, total cost, operating cost, etc.);
- Operational control (generation control, frequency deviation, reliability, etc.);
- Cyber security;
- Economic dispatch;
- Fault detection of equipment.

For ML methods, it is important that data from energy sources and consumers can be captured in large datasets, and that the data can be used for engineering appropriate EM systems. In the context of EM for buildings, privacy preserving methods are also researched [33].

The application of RL can be approached in two different ways, namely single and multi-agent RL [34].

2.3.1. Single Agent Reinforcement Learning

Qin et al. [35] proposed an approach to protect privacy of the load controls in residential microgrid. In their study, there is a central operator controlling a number of smart homes, and the authors suggest using Deep Reinforcement Learning (DRL) to solve the privacy issues. Muriithi and Chowdhury [31] used a PV microgrid including PV producer, Battery Energy Storage System (BESS), and local loads. It uses single agent RL algorithm Q-Learning to control the discrete charging and discharging behaviour of a battery. Ji et al. [36] used a microgrid consisting of distributed generators, PV installation, BESS, wind turbines, local loads, and a connection to the main grid. Here, a deep Q-network algorithm is used to train the distributed generators to produce cost-efficient energy. Khawaja Haider et al. [37] investigated in their study the differences of online and offline learning with respect to energy-storage systems in microgrids.

2.3.2. Multi Agent Reinforcement Learning

Samadi et al. [38] used a microgrid consisting of wind and PV energy sources, BESS, heat producers, diesel generators as a backup, and some thermal and electrical loads. This microgrid is connected to the main grid and represents each producer or consumer as an agent responsible for selling or buying energy in the EM market. The EM Performance is guided by reducing costs for the energy consumers in the microgrid. In this approach, Q-Learning is used, and the RE producers only contribute in the observation space but are not controllable. Foruzan et al. [39] used the RE microgrid, representing each component as an agent, RE producers as sellers, consumers as buyers, BESS as both, and the main grid as the seller. It uses an auction-based approach to maximise the profit within the microgrid using the Q-learning algorithm. Fang et al. [40] presented another residential microgrid focusing on an auction-based approach to achieve Nash-equilibrium within the microgrid using Q-learning algorithm. They consider residential PV producers as sellers and electric vehicles as storage acting as buyers. Fang et al. [41] proposed another Q-network approach to achieve Nash-Equilibrium in a residential microgrid. Here, the microgrid consists of a PV panel, wind turbines, distributed BESS, and industrial and residential loads. A game-theoretical view on multi-agent RL has been researched to find a point where it becomes uninteresting for an attacker to continue their attack [42].

2.4. DRL for Reliability and Resilience

To enhance the reliability of the Energy Management System (EMS) with DRL a multi-microgrid architecture where EM performance is guided by lowering the economic cost

of running, a RE microgrid has been proposed by Ref. [43]. In their model, four similar functioning microgrids are used, consisting of PV producer, wind turbines, BESS, diesel generators, and loads. Here, each microgrid has a microgrid control centre (MGCC), and there is also a central microgrid cluster control centre (MGCCC), each acting as an agent, with the main goal of working together and reduce the operational cost while performing load balancing.

To handle resilience and the impact of extreme weather events, a multi microgrid formation using a single agent RL to model a Distribution Network Operator (DSO) has been proposed [44]. The DSO uses network topology, total production, and total load as observation. The actions consist of changing the network configuration, while the reward is governed by voltage and current flow in a branch. Researching Multi-Agent Reinforcement Learning (MARL) and resilience, a double agent approach in a microgrid has been proposed [45]. One agent is a distributed generator while the other is a load agent and the EM performance is guided by optimal power flow. The optimal power flow is the total cost incurred by the generators to produce enough energy to achieve load balancing. The EM performance is also guided by improving the utility value of a microgrid in a short period of time with limited generation resources. As per Ref. [45], the utility value is given by load supply income, planned and unplanned outage losses, maximizing the utility value increases resilience. In another approach, a multi-agent framework has been used to enhance the resilience of the complete power system [46].

Q-learning is an important approach used to solve EM problems using RL,DRL or MARL. The usage of policy gradient methods with these actor-critic models is very interesting, as these have shown significant performance in teaching flexible movement to human simulation or walking to Boston Dynamics robot simulation [47]. In many of the studies there are simple rewards schemes as the EM performance is more guided towards profits and cost reduction. When the EM performance goal becomes multi-dimensional, the design of the reward scheme becomes complicated and the impact of it on the performance needs to be investigated.

3. Approach

In this section, we will describe the problem and the approach taken. Overall, not only in energy systems research but also, for example, in manufacturing, the number of data-driven approaches [48,49], in contrast to a model-driven approach [50], is increasing. We are following this trend.

Section 3.2 defines the problem statement and describes in brief the methodology mentioned in Figure 1. Section 3.3 describes the energy profile data, the data analysis techniques applied, and the generation of artificial data to promote the generalisation behaviour for trained agents. Section 3.4 discusses the MARL approach towards training an energy management simulation. Section 3.5 describes the approach towards qualitative and quantitative analysis of the trained agents.

3.1. Research Questions

Based on the above motivation, we define our research questions as follows:

- Is it possible to provide a tool to support decision-making with respect to the design and management of RE microgrids that allows to assess the reliability and resilience of microgrids?
- Is it feasible to develop a data-driven EM simulation for a RE micro grid, that leverages real-world energy data?
- To what extent does the simulated EM system show reliability and resilience?

3.2. System Design

To model real world problems, an EMS simulation must be developed where the RE producers, consumers, and a battery storage can be modelled as agents working together.

In this research, an EMS system is considered as a RE PV microgrid that includes the following components:

- Profile-Following PV producer;
- Profile-Following consumer;
- BESS;
- Fully Controllable Producer (FCP), such as Diesel generators, etc.;
- Freely Acting Consumer (FAC), such as pumped water storage, etc.

Profile-following agents are an important aspect of this simulation. Here, any real world time series data can be used. The profile in this EMS is a series of 96 time steps of energy production or consumption, where each time step is 15 min of a day. For our research, we are using energy profiles from Upper Austria large PV installations and industrial consumers.

The goal for the agents is EMS performance, which can be defined as:

- Profile Following: The profile-following agents should closely follow the energy profile defined for them;
- Battery Function: A BESS agent should be able to perform its duties of storing access RE and providing energy when the produced RE is not reaching the required consumption in the microgrid;
- Load Balancing: The FCP and FAC agents should act in a way so that energy is closely balanced in this microgrid;
- Control Energy or Load Balancing: The stochastic nature of RE and consumption introduces sudden fluctuations, where the BESS is not capable of reacting fast enough. This points out the need for a component of energy that can handle such changes of energy production and consumption. In this microgrid, FCP and FAC should provide this control or load balancing. FCP should provide the minimum energy required in the microgrid to balance energy while FAC should store the minimum of the excess energy generated in the microgrid.

This EMS performance is subject to weather conditions on different days of the year as the energy profiles on a summer day and a winter day will be quite different.

Finally, the impact of sudden drop in PV production due to weather conditions also needs to be studied. An analysis of EMS performance can be carried out, in the sense that in a more resilient microgrid each actor in the microgrid is able to react to sudden fluctuations, where the agent should still perform their duties. In a less resilient microgrid EMS performance will lead to a gap between balanced consumption and production, or agents will leave their profile.

A methodology, as described in Figure 1, has been developed to perform resilience and reliability analysis of a PV microgrid.

In brief, the methodology is as follows,

- The raw energy data, i.e., PV profile data and consumption profile data, are visualised using the tool mentioned in Appendix A to select a set of profile data;
- Signal analysis is performed on the data to generate synthetic profiles to avoid overfitting to specific profile data of a day;
- Once the synthetic profile is ready, energy agents in the energy management simulation are trained to achieve energy management performance;
- After this step, the trained agents are subjected to quantitative, reliability, and resilience analysis.

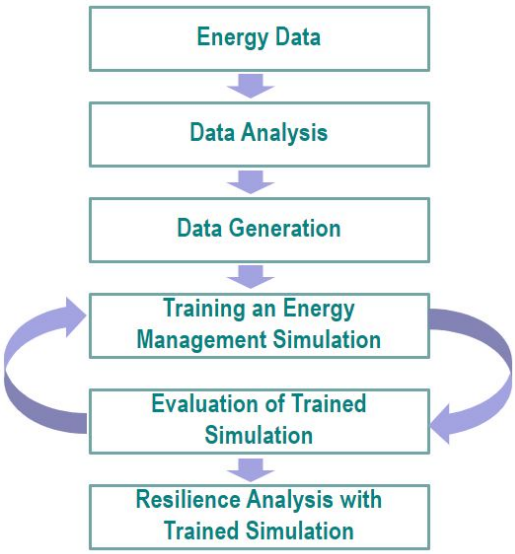


Figure 1. Proposed methodology.

3.3. Energy Data

This subsection gives an overview of the energy profile data used in this research. Furthermore, the data analysis techniques and synthetic data generation techniques are outlined in this section. This analysis of the data and synthetic data generation plays an important role in our reliability analysis strategy, where the agents are subjected to different types of profiles that capture the essence of different seasons for PV data or different industry work hours for industrial consumers.

3.3.1. Energy Profile Data

The training data for the EM simulation comprise of real-world energy profiles for large-scale PV installations and industrial consumers in Austria. Figure 2 shows sample profiles covering one work week in June 2019. The seventh day is Sunday and the factory is closed. This is visible by the consumer profile (blue). The dome shaped PV production production profiles (orange) show strong fluctuations on the 4th and 6th day, most likely due to variations in cloud cover on these days. Figure 2 is extracted using the tool developed to understand the data, the tool developed has been explained in Appendix A.

For training, data from a year is available. There are clear differences between winter and summer. The quality of the data and profiles varies. It is possible to explain observed peeks or valleys—apart from in some cases. In some cases sensors or data streams must be broken or interrupted.

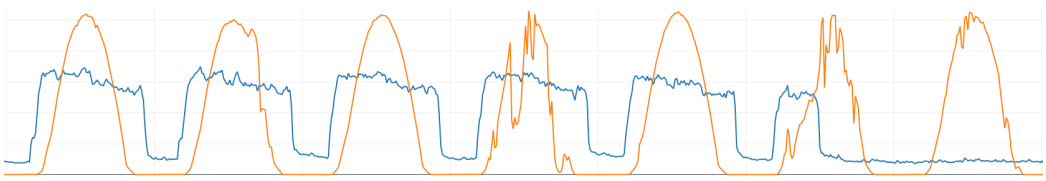


Figure 2. Data samples for the industrial energy consumption profiles (blue) and dome shaped PV production profiles (orange).

From the data of large scale PV installation, one month data for June 2019 is chosen as the final dataset for the experiments. This dataset is chosen as it contains different patterns of PV and consumption profiles. The maximum PV production is 280 kW and the maximum consumption is 180 kW in this dataset. For experiments, the energy values are scaled down by a factor of 10^5, and each profile has a data of 96 time steps comprising of

the whole day. This dataset of 30 days is later referred to as the energy profile dataset in the text.

3.3.2. Data Analysis

Data comprehension was achieved by a short analysis of the given profiles. To identify the underlying patterns isolated from the seasons, one month of the dataset was picked. As all profiles are time series, a trend and noise analysis showed potential classifications of PV and consumer behaviours. We compare the following analysis with Figure 2.

PV profiles can be described by taking the theoretical bell shape, resulting from cloudless and perfect conditions, as a baseline. This baseline is dependent on the season, as amplitude is varying over the time of a year. We were able to identify three different cases of perturbations. First, the baseline is nearly achieved, with negligible perturbations and up to 5% amplitude variation from the median. Second, 5–7 short intervals of 2–3 time steps, with a low impact up to 20% change relative to the baseline. Third, longer intervals with higher impact and shocks up to 60% change relative to the baseline. In the second and especially in the third case, daily variations result in strong noise throughout the day.

Consumer profiles can be described by summing weekday dependent needs to the minimum consumption, as baselines. This results in three different cases. Minimum consumption varies consistently between 0.1 and 0.4 power units. Workdays require an additional average consumption of 1.5 power units starting in the first quarter of the day. Dependent on full or half workdays, the power is needed until the end or the middle of the day. Daily variations result in noise throughout the day.

3.3.3. Data Generation

After reviewing the results from Section 3.3.2, generation of artificial training data for PV and consumer profiles can be achieved. Our strategy enables for new profiles, mimicking the behaviour of PV and consumer profiles in the dataset. Following the dataset ratios, a case combination is randomly picked from the previously described possibilities.

To generate artificial PV profiles, we use a relatively simple multiplicative time series approach. The time series consists of two parts: the above mentioned baseline and a noise. Dependent on the chosen class, the baseline is adapted and the noise model varies. In class 1, only slight deviations from the baseline were observed. For simplicity, we decided to generate each noise array from an uniformly distributed random number generator with a small range of 0.95–1.0. Classes 2 and 3 need more sophistication. However, for both we apply the same kind of noise array at last to imitate small daily fluctuations and add broadness to the data. In class 2, the noise model follows the observations and decides randomly upon which and how many intervals in the time series to pick for stronger perturbation. Given the intervals, impacts are randomly decided on, with a maximum of 20% deviation from the perturbed baseline. In each interval the impact is applied at one random chosen timestamp and the neighbouring interval timestamps are interpolated, following polynomial interpolation of second degree. A similar approach, consisting of more and more impacting intervals, is used in class 3 right before the last noise array of class 1 style is applied and after the step explained next, to imitate the original dataset. From observations of the data, class 3 profiles can contain a shock or dent of higher impact, which has to be modelled separate of the interval approach, as they appear once or twice and have a potentially longer duration. It is randomly chosen if and where such an event occurs in the time series: namely in the first half, the second half, both, or overall. Based on this, if chosen, the baseline is reshaped by application of 50–60% value reduction at 1 to 30 neighbouring points in the given area. By interpolating, following the polynomial interpolation of second degree in a randomly chosen size of a neighbourhood, the new baseline shape is generated.

Consumer profiles are generated by applying a direct one-layer noise model to the case specific baseline. The noise model is generated from a standard normal distributed random number generator. Each baseline follows the observed consumption needs, where

the consumption levels and intervals are randomly chosen, based on the occurrences in a given case.

3.4. MARL Approach to Energy Management Simulation

This section gives an overview of the RL approach to model the problem stated in Section 3.2. It is based upon our earlier work [34] and extends it with generalisation analysis and inclusion of stressors in the simulation. Section 3.4.1 gives a brief introduction to RL and MARL, the simulation model is introduced in Sections 3.4.2 and 3.4.3 describes the training environment, observation space, DRL model, and the reward scheme used to steer the agents. Section 3.4.4 discusses the development accomplished to achieve generalisation of the trained MARL model. Finally, Section 3.4.5 introduces the definition of stressors used to stress the simulation.

3.4.1. RL and MARL

Deep Reinforcement Learning

RL is regards developing decision-making skills through sequential interactions with an environment. The environment defines an observation space $\mathcal{S}$ and an action space $\mathcal{A}$. At each time step in the interactions, the RL agent receives an observation s_t and a reward r_t from the environment and then decides on an action a_t by using a policy function $\pi(a_t|s_t)$. The learning objective for the agent is to maximize the expected cumulative reward:

$$R_t = \sum_{k=1}^{\infty} \gamma^k r_{t+k+1}, \quad \gamma \in (0, 1]. \tag{1}$$

The definition of the value of a policy π for a state s is

$$v_\pi(s) = \mathbb{E}_\pi(R_t|s_t = s), \tag{2}$$

and the action-value function of a policy is defined as

$$q_\pi(s, a) = E_\pi(R_t|s_t = s, a_t = a) \tag{3}$$

when an action a is taken in a state s. Maximisation of the action-value function results in an optimal policy.

With the most recent developments in Deep Learning (DL), several new opportunities in Machine Learning have emerged. The combination of DL with RL (DRL) has in particular produced new, astonishing outcomes in a variety of sectors, such as superhuman performance in video games. In DRL, deep neural networks are used as function approximators for value and policy functions. This introduces a network parameter θ, which allows to directly optimise the policy by looking for the best values for θ in the policy space $\{\pi_\theta(a_t|s_t), \theta\}$. For bigger spaces of states and/or actions where a tabular representation is impractical, function approximation is essential.

The gradient ascent approach (Baird and Moore [51]) can be used to optimise the neural network parameters θ, leading to a class of algorithms known as policy gradient methods. A parameter update is represented by an estimate in the gradient of an objective function. Proposed by Schulman et al. [52], the following objective function is frequently employed for policy gradient methods:

$$L^{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t] \tag{4}$$

where $\hat{A}_t$ is an estimator of the advantage function that describes the additional benefit that could be gained by acting in the manner indicated by a_t.

So-called actor-critic methods are created by the combination of policy gradient methods with action-value functions. The critic approximates the action-value function while the actor approximates the policy, criticising the actions that the policy has taken.

Multi-Agent Reinforcement Learning

MARL is a multi-agent generalisation of RL, which studies how multiple agents learn in a shared environment. The presence of additional agents, who are actively changing the training environment, presents a significant difficulty for MARL. An agent's observation includes information on both the agent's own activities and those of other agents. In other words, several agents in MARL indirectly communicate with one another through their behaviour. Dependencies and load balancing in the EM scenario necessitates coordinated activity from all agents. We took the Reinforcement Learning library (RLlib) implementation of a Proximal Policy Optimisation (PPO) algorithm and enhanced it with a centralized critic to be able to train coordinated actions. Inspiration for applying a centralised critic approach is taken from Yu et al. [53].

The new family of actor-critic approaches, known as PPO, was put forth by Schulman et al. [52]. A concept to stabilize training was suggested using a modification of (4). Large policy changes are constrained by the new objective, which results in smaller steps and allows for numerous epochs of mini-batch updates. The new aim is established as follows using the ratio of the new and old policies: $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t))$,

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t] \quad (5)$$

where $clip(r_t(\theta), 1-\epsilon, 1+\epsilon)$ clips the ratio to the interval $[1-\epsilon, 1+\epsilon]$.

3.4.2. Simulation

As per Section 3.2, the components of RE PV microgrid are considered in the simulation as agents that need to be trained to maintain EM performance. As with any RL problem, the development of an efficient and effective training environment is crucial. This essentially boils down to the design and implementation of agents' observations, actions, and rewards. To capture EM dynamics, the observations in the simulation training environment consist of time series for four quantities and these four quantities depend on the type of agents. For profile-following agents and fully controllable agents, this is the total energy production in the microgrid, total energy consumption, the actual load of the agent, and the load according to the agent's energy profile. For a BESS agent, it is the total energy production in the microgrid, total energy consumption, current battery State Of Charge (SOC), and total renewable energy available. At any time step, an agent can choose 1 of 11 discrete actions: increase load (5 discrete increments), decrease load (5 discrete decrements) or do nothing. The EM performance forms the basis for the reward scheme design of the MARL based simulation.

3.4.3. Agent and Training Environment

The five components specified in Section 3.4.2 are represented as agents in the multi-agent compatible RL training environment that is described in this subsection. The goal of the training is to discover the best strategies for achieving the objectives of energy management outlined in Section 3.4.2.

- **Training Environment**:
  The multi-agent environment of RLlib [54] has been bootstrapped for the creation of the training environment, making it compatible with OpenAI gym environments. Box observation spaces and discrete action spaces are used in the environment for the agents. Tang and Agrawal [55], who claim that "the discrete policy provides considerable performance advantages with state-of-the-art on-policy optimization methods PPO" served as an inspiration for the choice to employ discrete action spaces with the PPO algorithm. The optimal number of discrete sampling for a continuous action space is given by Tang and Agrawal in their citation of Ref. [55] as being (7–15). In our studies, 11 discrete actions produced the best results;

- **Parameters and Variables**:
  In the Tables 1 and 2, the mathematical notation and the variables for the EM problem are introduced, which will be used throughout the paper;
- **Agents**:
  The characteristics of the agent configuration are detailed below. For profile-driven agents, the energy profile is the decisive aspect, whereas for BESS, their initial SOC and minimum SOC play a crucial role. In FCP and FAC, load balancing tolerance is a critical part. Load-balancing is accomplished if the absolute difference between total production and total consumption is less than the tolerance. Max-load-diff is the maximum load difference between two consecutive time steps for each agent. In BESS it establishes the maximum reaction magnitude. As shown in Figure 3 the observation of an agent is made up of four time series with five time steps each.

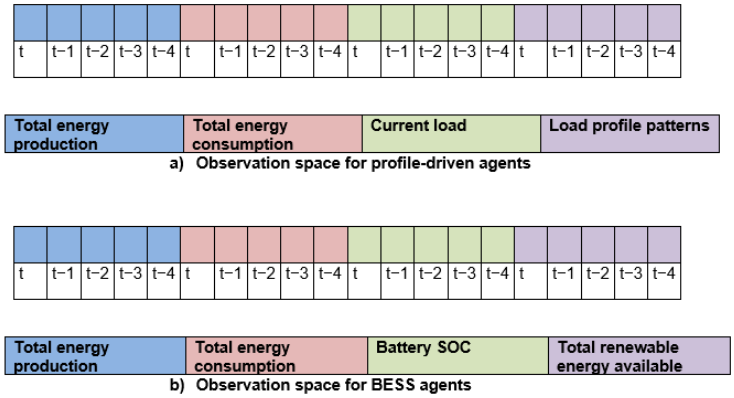


Figure 3. Observation space for agents.

Parameters

Table 1. Mathematical notations for the parameters.

Parameters	Explanation
T	The number of time steps per episode.
SOC_{init}	Initial SOC value of the BESS.
SOC_{min}	Minimum SOC value of the BESS.
SOC_{max}	Maximum SOC value of the BESS.
P_{max}	Maximum load of the PV producer.
F_{max}	Maximum load of the fully controllable producer.
C_{max}	Maximum load of the profile-driven consumer.
D_{max}	Maximum load of the freely acting consumer.
B_{max}	Maximum battery magnitude of the BESS.
Δ^{pv}	Maximum load difference of the PV producer.
Δ^{f}	Maximum load difference of the fully controllable producer.
Δ^{c}	Maximum load difference of the profile-driven consumer.
Δ^{d}	Maximum load difference of the freely acting consumer.
F_{init}	Initial load of the fully controllable producer.
D_{init}	Initial load of the freely acting consumer.
R_t^{pv}	Profile load for PV producer at time t, $t < T$.
R_t^{c}	Profile load for profile-driven consumer at time t, $t < T$.

Variables

Table 2. Variables mathematical notations.

Variables	Explanation
p_t	Load of the \ac{pv} producer at time t, $t < T, 0 \leq p_t \leq P_{max}$.
f_t	Load of the fully controllable producer at time t, $t < T, 0 \leq f_t \leq F_{max}$.
c_t	Load of the profile-driven consumer at time t, $t < T, 0 \leq c_t \leq C_{max}$.
d_t	Load of the freely acting consumer at time t, $t < T, 0 \leq d_t \leq D_{max}$.
b_t	Battery magnitude of \ac{bess} at time t, $t < T, -B_{max} \leq b_t \leq B_{max}$.
soc_t	\ac{soc} of \ac{bess} at time t, $t < T, SOC_{min} \leq soc_t \leq SOC_{max}$.

In the following, the agents are specified in more detail.

- PV producer agent:
  The energy output profiles of a PV panel are followed by this agent, thus the agent is profile-driven. Figure 3a shows the agent's 20-dimensional observation space. The agent has a discrete action space of 11 non-negative numbers, with the options (0–4) for reducing production load, 5 for doing nothing, and (6–10) for increasing production load. The production load increase or decrease is thus represented as $(0.2, 0.4, 0.6, 0.8, 1.0) * \Delta^{pv}$;
- Profile-driven consumer agent:
  This agent adheres to power consumption profiles and is profile-driven. As with the PV agent, the observation and action space are identical;
- BESS agent:
  This agent mimics the behaviour of a battery storage, whose primary goal is to charge and discharge batteries in an acceptable manner, which is to charge when renewable energy is present, and discharge when no or less renewable energy is present. Figure 3b illustrates its 20-dimensional observation space. Its action space is made up of 11 non-negative values, where 0–4 corresponds to battery drain, 5 to inactivity, and 6–10 to battery charging. Max-load-diff, in the context of the BESS agent refers to a battery's maximum rate of charging and discharging B_{max}. The battery magnitude, also known as the effective charging/discharging rate, is calculated as $(0.2, 0.4, 0.6, 0.8, 1.0) * B_{max}$. Specific configuration parameters for the BESS agent are initial/minimum/maximum SOC levels, denoted as SOC_{init}, SOC_{min} and SOC_{max}. The initial value refers to the start value at the beginning of an EM episode;
- FCP and FAC agent: Both agents share their specifications for observation space and action space with the PV agent. The maximum power output that can be sent into the microgrid is a crucial configuration factor for the fully controllable producer. These agents have no specific profile to follow and are mainly in charge of load balancing.

- **Deep Reinforcement Learning Model**:
  In our implementation, which builds upon RLlib's implementation of a PPO algorithm, the agents share a centralised critic model, criticising the agents' actions from a microgrid-wide perspective. The DRL model that was utilized for each agent is depicted in Figure 4. The action logits are located in the third layer of the actor model's three layers. Each agent's centralized critic model comprises three input levels. Let us assume that we have n agents in order to better comprehend the input layers. The first input layer is the agent's own observation with the shape $(, 20)$; the second input layer processes the observations of the opponent agents with the shape $(, 20 * (n-1))$; and the third layer processes the actions of the opponent agents with the shape $(, 11 * (n-1))$.

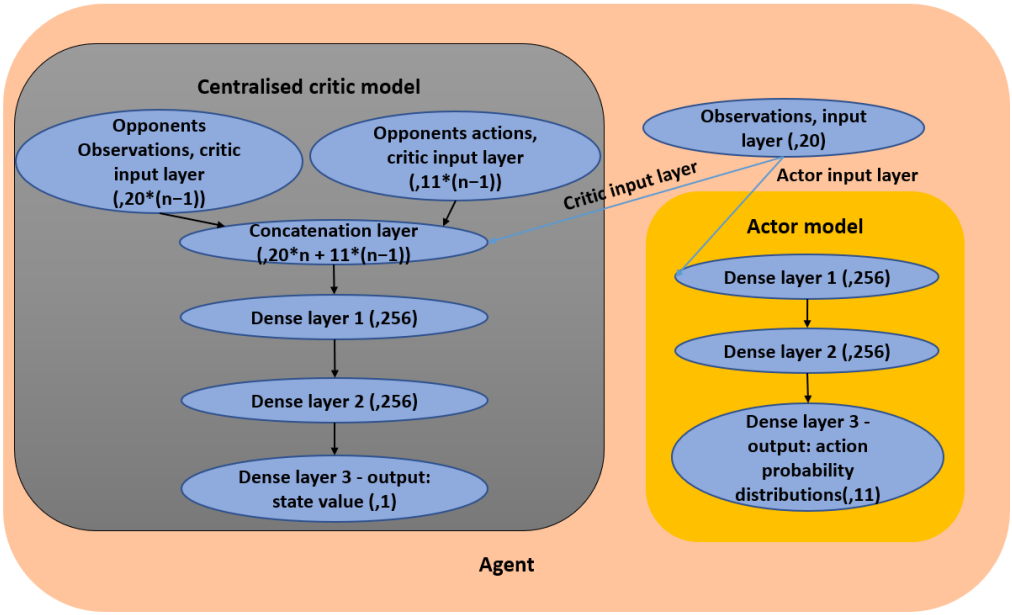


Figure 4. DRL model.

The three input layers are concatenated, and then the combined input is processed by two dense hidden layers. The last layer generates a single value, which is the action-value for a specific input of observations and actions. The first dimension of $(,\text{size})$ for all layers is left unspecified because it depends on the (configurable) mini-batch size of the PPO algorithm.

- **Reward Scheme for Agents**:
  The reward scheme for different agents in our EM simulation is as shown in Table 3.

Table 3. Agents and reward schemes.

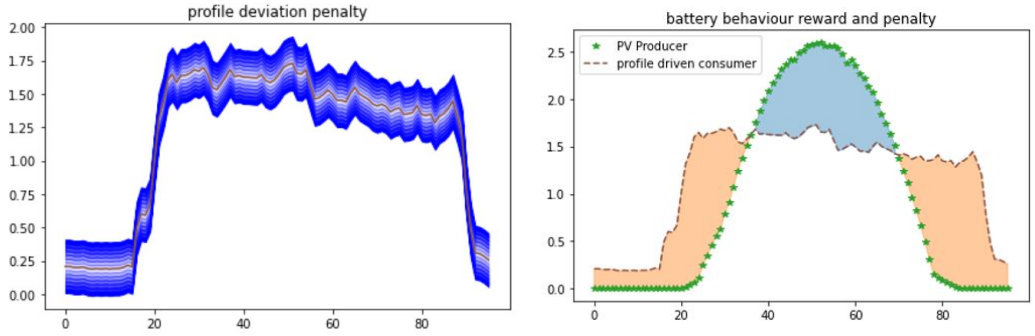
Agent	Profile Deviation Penalty	Load Balancing Reward	Battery Behaviour Reward and Penalty	Excess Energy Penalty and Appropriate Energy Reward
PV agent	X			
Profile-Driven Consumer	X			
BESS			X	
FCP		X		X
FAC		X		X

In the following, we discuss what these different reward configurations mean. There are four types of reward configuration as mentioned below:

- Profile Deviation Penalty: As displayed in Figure 5a, the profile-following agents should remain close to the energy profile to avoid distance-based growing penalty. In Figure 5a, shades of blue extend to infinity in both directions, showing distance-based penalty;
- Battery behaviour rewards and penalty: Shown in Figure 5b BESS should provide energy in the orange area and store the RE in the blue area to avoid penalty and earn rewards;
- Load Balancing Reward: FCP and FAC agents are given a reward if they work together such that (total production − total consumption $\leq$ balance tolerance) where balance tolerance is one of the environment configurations;

- Excess energy penalty and appropriate energy reward: As per Figure 5b there are different additional reward configurations for FCP and FAC agents;
  - FCP: It receives a reward for producing appropriate energy in an orange area and producing nearly zero energy in a blue area, otherwise, a penalty is applied;
  - FAC: It receives a reward for absorbing nearly zero energy in an orange area and absorbing appropriate energy in a blue area, otherwise, a penalty is applied.

Complete equations to reproduce these reward schemes can be found in the work by Haemmerle et al. [34].



(**a**) (**b**)

Figure 5. Reward scheme in a nutshell. (**a**) Profile deviation penalty. (**b**) Battery behaviour reward and penalty.

3.4.4. Generalised Training

This subsection talks about the development accomplished to enable generalisation capabilities of the trained agents. Data analysis from Section 3.3.2 and artificial data generation from Section 3.3.3 form the basis for this development. The PV classes mentioned in Section 3.3.3 represent the energy profiles of different days taken from energy data. During training, based on the randomly generated profiles from a specific PV class for producer and consumer class for consumer, the profiles are exchanged after every episode. This helps the trained agents to react to fluctuations that may occur for a specific PV class or a consumer class. For evaluation, a specific day is picked from the energy profile dataset corresponding to the specific PV class and consumer class. Then a complete quantitative analysis is carried out based on Section 3.5.1.

3.4.5. Inclusion of Stressors in the Training Environment

To enable Resilience and Robustness Analysis, we included the option to use abstract stressors in the environment. To simplify the approach, we only focused on PV and consumer agents, while using a rather straightforward stress signal. At any given sub-interval of an episode, one of the two agents can be manipulated. Stress is referring to a change in the produced energy for PV or consumed energy for the consumer. Whereas the profile-following is not influenced, the observation from other agents will be. Especially the BESS, FCP and FAC agents are dependent on the profile-following agents and need to adapt to the new situation.

The parameters for stressors are determined by their interval and significance. Inspired by real-world scenarios, we decided upon different levels of impact. Low, moderate, and high impact, representing high, moderate, and low probability events.

3.5. Simulation Analysis

For simulation analysis, one or multiple days from the dataset are selected. The analysis is carried out using the following three steps:

- Evaluation with no stressors: Evaluation against original profiles from selected data is performed. Resulting plots and quantitative analysis as per Section 3.5.1 are recorded;
- Evaluation with theoretical optimum: Evaluation against the theoretical optimum simulation as defined in sub-Section 3.5.2 is carried out on selected data;
- Evaluation with stressors switched on: Trained simulation is perturbed with stressed PV profile data or consumption data or both, which is described in sub-Section 3.5.3 for selected data. The resulting plots and quantitative analysis are recorded.

3.5.1. Quantitative Analysis

This subsection describes the quantitative analysis formulation used in this article. Below are the legends for the quantitative analysis of the simulation:

- mae—Mean Absolute Error (MAE);
- rmse—Root Mean Squared Error (RMSE);
- norm rmse/mae—normalized ratio RMSE against MAE;
- sma—Simple Moving Average;
- sma mae—MAE between SMA's;
- noice—additive signal decomposition in SMA's and noise;
- noice metrics—statistical noise signal analysis based on types;
- ee—excessive energy produced, considering PV producer, profile-driven consumer and BESS;
- EE FCP—total excessive FCP production, while the PV producer exceeds the profile-driven consumer, i.e., energy should have been provided by the renewable producer but still the non-renewable produces the energy;
- AEE FCP—total excessive energy available higher than the balance tolerance;
- diff AEE FCP—total absolute FCP excessive energy higher than ee. For example FCP energy is 0.75 units and ee is 1.5, then diff AEE fcp would be 0.75, similarly if FCP energy is 1.5 and ee is 0.75, diff AEE, fcp would again be 0.75;
- EE FAC—total excessive consumption, while PV production is less than profile-driven consumption, i.e., total RE available is less than consumption;
- diff AEE FAC—total FAC excessive energy absorbed higher than ee. For example FAC energy is 0.75 units and ee is 1.5, then diff AEE fac would be 0.75, similarly if FAC energy is 1.5 and ee is 0.75, diff AEE fac would again be 0.75;
- Storage illegal actions—number of time steps battery not charging or discharging appropriately.This means the battery is charging when no or less renewable energy is present or the battery is discharging when there is no consumption available;
- Storage absolute illegal loads—total battery magnitude difference to the difference of RE—consumption available.

The quantitative analysis can be divided into four parts:

- Profile-Following: This is quantitatively analysed based upon how close the agents are following the energy profile. Distance over an episode is measured through mae, while rmse and norm rmse/mae give more insight to the occurred spikes. Sma mae and noise metrics on the other side analyse from a behavioural standpoint. Small sma mae points towards a good underlying baseline and low-level understanding of noise trends. Noise metrics investigates if the high-level noise model is matching. These measures provide a deeper insight into the comparison of the energy profile and the solution of the trained agent;
- Storage Behaviour: for storage to behave appropriately it is important that it produces energy when no or less RE is available, while it absorbs energy when there is excess. To analyse this behaviour quantitatively, the storage illegal actions and storage absolute illegal loads are used;
- Load Balancing: For load balancing, the distance between total production and total consumption is measured. As above, the metrics are mae, rmse, and norm rmse/mae;
- Control Energy: In addition to load balancing, FCP and FAC agents can be quantitatively analysed by measuring how little energy is produced by the FCP agent and

how little energy is absorbed by the FAC agents. For FCP agents EE FCP, AEE FCP, and diff AEE FCP provide the analysis while for an FAC agent it is EE FAC and diff AEE FAC.

The above quantitative analysis gives an analysis towards the energy management simulation. Furthermore a comparison between single-day trained agents and generalised trained agents can be carried out by keeping the same metric in mind.

3.5.2. Comparison to Theoretical Optimum

As an additional component for quantitative evaluation of trained agents, a linear program for the EM problem described in Section 3.2 has been implemented. The input parameters and variables for the linear program are described in Section 3.4.3. The linear program solves the EM problem for one day, and it requires the energy profiles to be fully specified for the day under consideration. This renders the mathematical programming approach impractical for deployment scenarios, where EM control decisions have to be made sequentially, and future consumer loads, as well as future energy output from PV producers, are not known. However, a multi-agent system trained with RL is able to cope with deployment scenarios, because the agents' control decisions are based on their observations, and these do not contain any information about the future.

The linear program has been implemented with CMPL 1.11.0 (<Coliop | Coin> Mathematical Programming Language). In the following, the objective function is briefly discussed.

$$O = \min \sum_{t<T} \Big\{ W^{pf} \cdot \Big(\Big| p_t - R_t^{pv} \Big| + |c_t - R_t^c| \Big) + W^{lb} \cdot |p_t + f_t - b_t - c_t - d_t| + W^{ee} \cdot (f_t + d_t) \Big\} \tag{6}$$

In Equation (6) the parameters W^{pf}, W^{lb}, W^{ee} denote the weights for the profile-following, load balancing, and excess energy term, respectively. If $b_t < 0$, the BESS is discharging (acting as producer), and if $b_t > 0$, the BESS is charging (acting as consumer).

The linear program provides exact solutions to EM problem instances, and these solutions are used as benchmarks for the trained simulation. For benchmarking, an integrated software is developed that uses CMPL's Java API, together with CMPL's built-in Cbc (Coin-or branch and cut) solver 2.9.8 to calculate the benchmark solutions. Comparison of each agent with a benchmark is carried out considering their 96 time steps of load or battery SOC data. All the loads or battery SOC of trained agents and benchmark solution, respectively, are compared using the profile-following metrics, as mentioned in Section 3.5.1. So, we compare the load or battery SOC of each agent (benchmark vs. trained) considering mae, rmse, norm rmse/mae, sma mae, and noise metrics. For load balancing comparison (benchmark vs. trained), we use the profile-following metrics to compare the difference (total production–total consumption) curves for each solution.

3.5.3. Reliability and Resilience Analysis

In our work we use the following definition of EM reliability: "EM in a microgrid is reliable, if it is able to provide high quality energy to the microgrid's energy consumers for a large variety of EM problem instances". Energy quality is a direct consequence of load balancing in the microgrid. With better load balancing the quality increases, hence load balancing mae values are used to measure reliability. Reliability calculation is formalised as follows, assuming that the inverse reliability value σ_M is specific to a set of EM problem instances M:

$$\sigma_M = \frac{\sum_{m \in M} l_m}{|M||} \tag{7}$$

In Equation (7), the load balancing mae value for problem instance m is denoted by l_m.

In this paper EM resilience is defined as a microgrid's ability to provide high quality energy to consumers in the microgrid, when the microgrid is confronted with stressors. For resilience analysis, the following stressors are used:

- Distinct indentations in PV energy output;
- Distinct indentations in energy consumption.

For a specific stressor s and a problem instance m, the inverse resilience value ρ_{sm} is defined as

$$\rho_{sm} = \frac{l_m^s - l_m}{l_m} \tag{8}$$

In Equation (8), l_m denotes the load balancing mae value for the undisturbed problem instance m, and l_m^s denotes the load balancing value for the stressed problem instance m.

4. Results

This section describes the evaluation results for trained EM simulations. In the following, the evaluation results are presented in two subsections: (1) single-day training and (2) generalised training. In each of these subsections, two cases are discussed: (a) evaluation without stressors, and (b) evaluation with stressors. In case (a), undisturbed energy profiles are used for evaluation. In case (b), the energy profiles are modulated by a stressor signal. The evaluation uses the formulation for quantitative analysis laid out in Section 3.5. Additionally these subsections discuss how the agents in the EM simulation are trained, and the EM problem instances that were used for training and evaluation.

4.1. Single-Day Training

For single-day training and evaluation, a subset of three days $\{1, 2, 3\}$ from the energy profile dataset is used. Each day comprises specific energy profiles for PV producer and profile-driven consumer, cf. Figure 6. However, the energy profiles are just part of a day-specific EM problem instance. The full specification of the three problem instances that are used for training and evaluation is shown in Table 4.

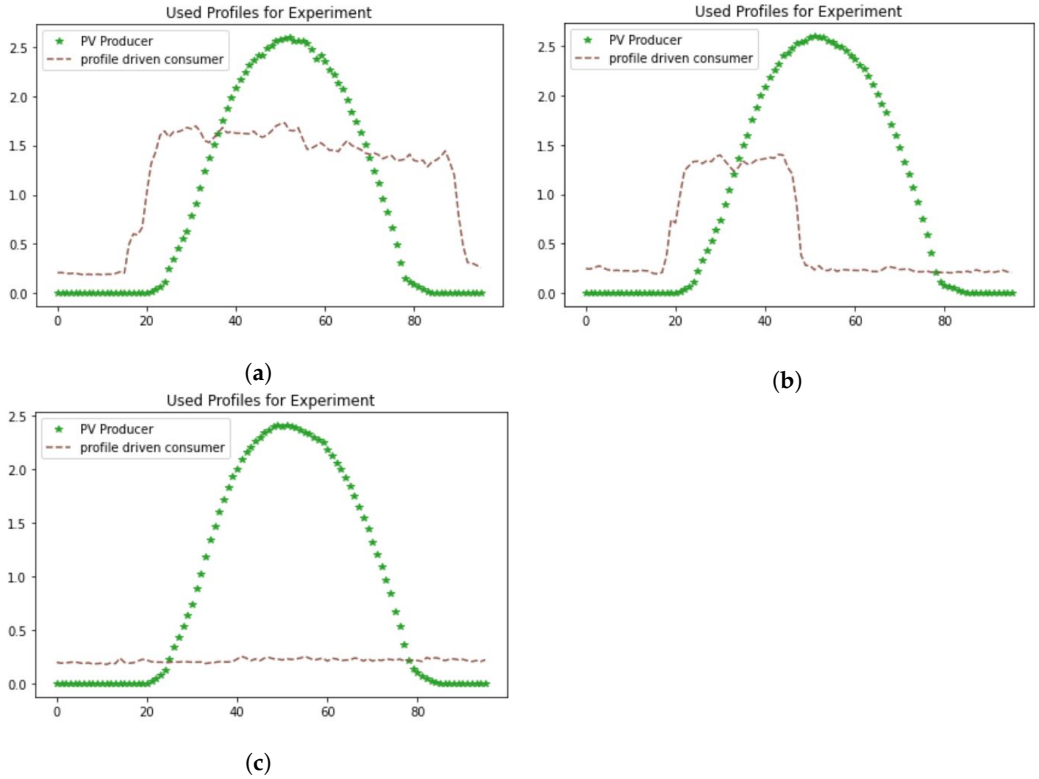


Figure 6. Energy profiles of the problem instances. (**a**) Day 1, (**b**) day 2, (**c**) day 3.

Table 4. Problem instances for EM simulation.

Parameters	Day 1	Day 2	Day 3
PV profile	Dataset day 1 Figure 6a	Dataset day 2 Figure 6b	Dataset day 3 Figure 6c
Consumption Profile	Dataset day 1 Figure 6a	Dataset day 2 Figure 6b	Dataset day 3 Figure 6c
SOC_{init}	30.0	30.0	30.0
SOC_{min}	2.4	2.4	2.4
SOC_{max}	60.0	60.0	60.0
P_{max}	5.0	5.0	5.0
F_{max}	5.0	5.0	5.0
C_{max}	5.0	5.0	5.0
D_{max}	5.0	5.0	5.0
B_{max}	1.5	1.5	1.5
Δ^{pv}	0.187510	0.194949	0.166984
Δ^{f}	0.5	0.5	0.5
Δ^{c}	0.416	0.524191	0.052
Δ^{d}	0.5	0.5	0.5
F_{init}	0.208	0248	0.2
D_{init}	0	0	0

4.1.1. Evaluation without Stressors

This section reports on the results with respect to evaluating the trained EM simulation with problem instances from Table 4. The energy profiles are undisturbed, i.e., they are not modulated by stressor signals. The results cover two important cases: (1) training and evaluation use the same problem instance, (2) training and evaluation use different problem instances. Obviously, case 2 tests the generalisation capability of the trained EM simulation.

The three problem instances that are used for evaluation differ mainly in their consumption profiles, cf. Figure 6. Day 1 represents a normal production day with high energy consumption, where all the solar energy can be used by the profile-driven consumer and the BESS. Days 2 and 3, representing production on a Friday or weekend day, respectively, show the need for negative control energy, implemented by the FAC agent.

As shown in Figures 7–9, the single-day trained agents show good EM performance when they are evaluated with the trained days. The agents' load curves are depicted in Figures 7a, 8a and 9a. On day 1, when there is no need for negative control energy, the FAC agent is not active. However, on day 2 and day 3 the FAC agent is actively providing the negative control energy required for load balancing. On all days the BESS agent shows the required behaviour of providing energy when needed and absorbing excess energy from the PV producer whenever needed. The FCP agent is active whenever there is too little output from the PV producer. The profile-driven agents follow their profiles closely. Most of the time the trained agents show good load balancing, depicted in Figures 7b, 8b and 9b. One exemption is the period 50–80 in Figure 9b, where load balancing is inappropriate. This is due to the fact that the maximum storage capacity is reached, and the FAC agent is not properly trained to handle such a situation. We will see later on in Section 4.2, that this situation is mitigated with generalised training.

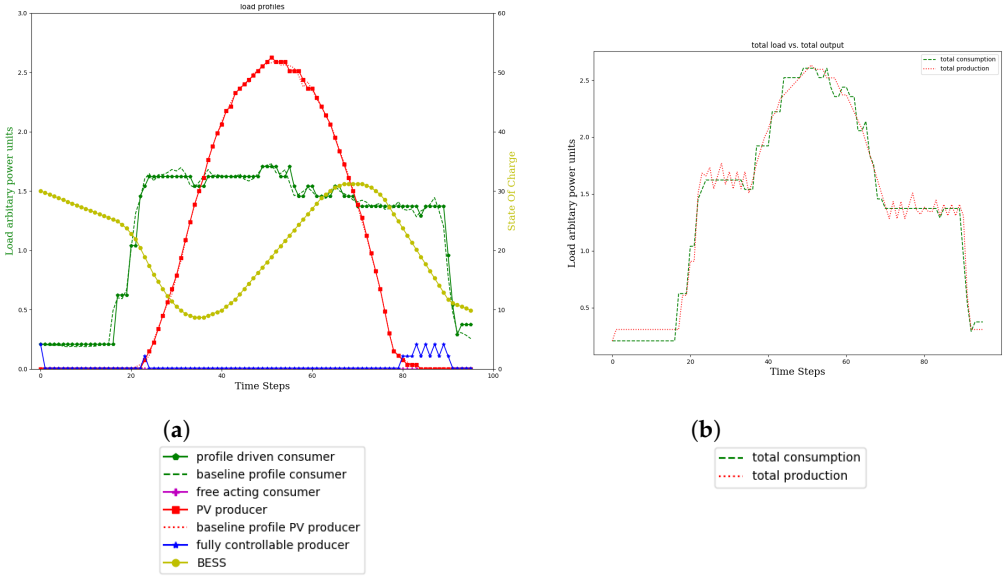


Figure 7. Evaluation on day 1, with agents trained on day 1. (**a**) Agent load curves and profiles. (**b**) Load balancing.

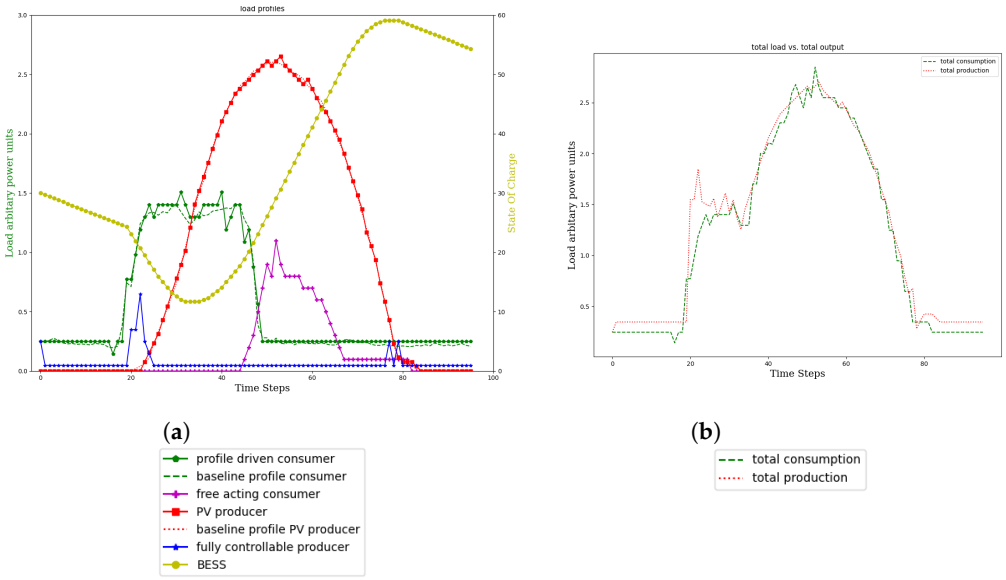


Figure 8. Evaluation on day 2, with agents trained on day 2. (**a**) Agent load curves and profiles. (**b**) Load balancing.

Figures 10 and 11 illustrate results for the case when training and evaluation use different problem instances. For both figures the agents have been trained on day 1. For Figure 10, the evaluation day is day 2, and for Figure 11 the EM simulation has been evaluated on day 3. In both figures poor load balancing is evident, due to the weak generalisation capability of the trained EM simulation. This is particularly obvious for the FAC agent. Being trained on day 1, the FAC agent never learned to provide negative control energy. However, the provision of negative control energy is a required behaviour for proper load balancing on days 2 and 3.

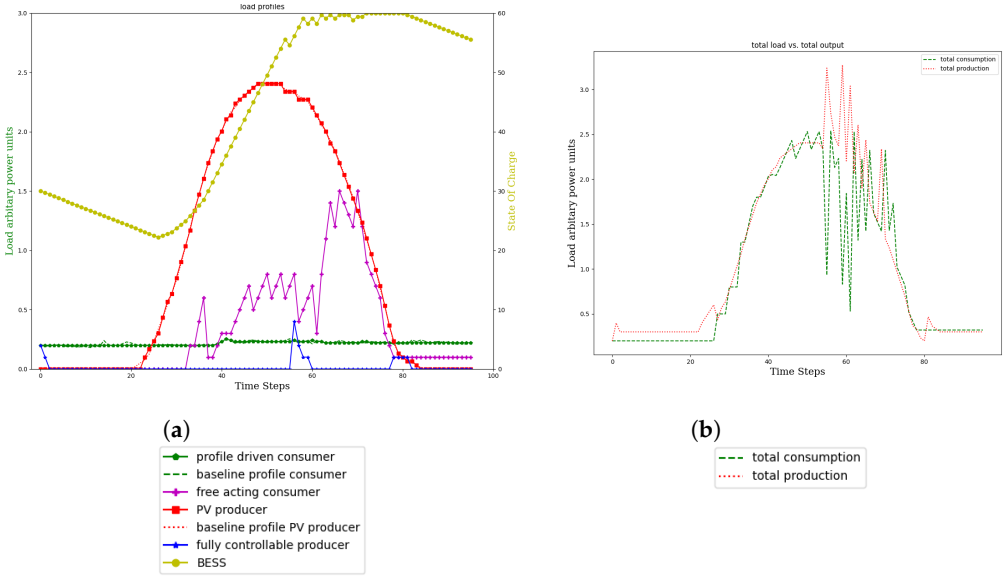


Figure 9. Evaluation on day 3, with agents trained on day 3. (**a**) Agent load curves and profiles. (**b**) Load balancing.

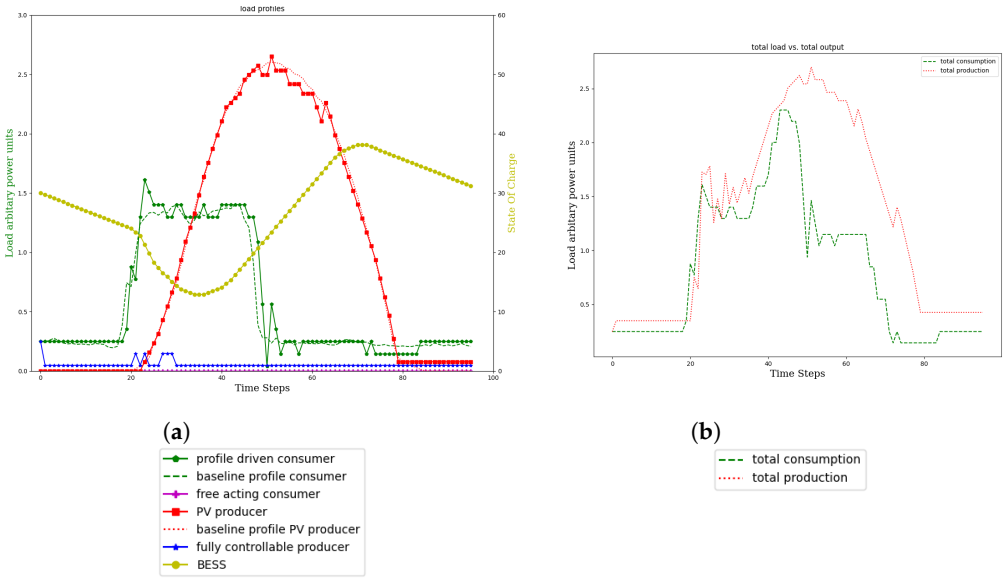


Figure 10. Evaluation on day 2, with agents trained on day 1. (**a**) Agent load curves and profiles. (**b**) Load balancing.

Table 5 shows quantitative analysis results of the trained agents, when evaluated on days other than the ones they were trained on. Table 5 points out an important drawback of single-day training: the inability to adapt to major changes in energy profiles. A profile-driven consumer agent trained on day 3 is unable to follow the consumption energy profiles of days 1 and 2, as these do include different patterns. Moreover, BESS and FAC behaviours are incorrect, resulting in poor load balancing when the single day trained agents are evaluated on different problem instances.

Table 5. Quantitative analysis of trained agents evaluated on different days.

Metric	Day 1 Training Day 1 Evaluation	Day 1 Training Day 2 Evaluation	Day 1 Training Day 3 Evaluation	Day 2 Training Day 1 Evaluation	Day 2 Training Day 2 Evaluation	Day 2 Training Day 3 Evaluation	Day 3 Training Day 1 Evaluation	Day 3 Training Day 2 Evaluation	Day 3 Training Day 3 Evaluation
Profile following PV (mae)	0.011	0.037	0.063	0.031	0.013	0.016	0.015	0.015	0.009
Profile following consumer (mae)	0.039	0.068	0.015	0.101	0.038	0.019	0.50	0.242	0.008
EE FCP	0.28	2.16	0.0	0.28	2.312	5.4	0.34	2.352	0.9
Diff AEE FCP	0.0	0.52	0.0	11.7	0.64	0.0	6.794	3.327	0.0
EE FAC	0.0	0.00	0.0	0.0	0.4	0.1	0.8	1.8	1.7
Diff AEE FAC	0.0	33.646	49.91	24.67	0.00	21.756	20.641	15.06	13.064
Storage illegal actions	0.0	7.0	16.0	5.0	0.00	4.0	8.0	7.0	7.0
Storage absolute illegal loads	0.0	6.296	11.382	3.042	0.00	2.337	12.041	11.325	16.08
Load Balancing (mae)	0.079	0.488	0.594	0.483	0.112	0.361	0.416	0.282	0.222

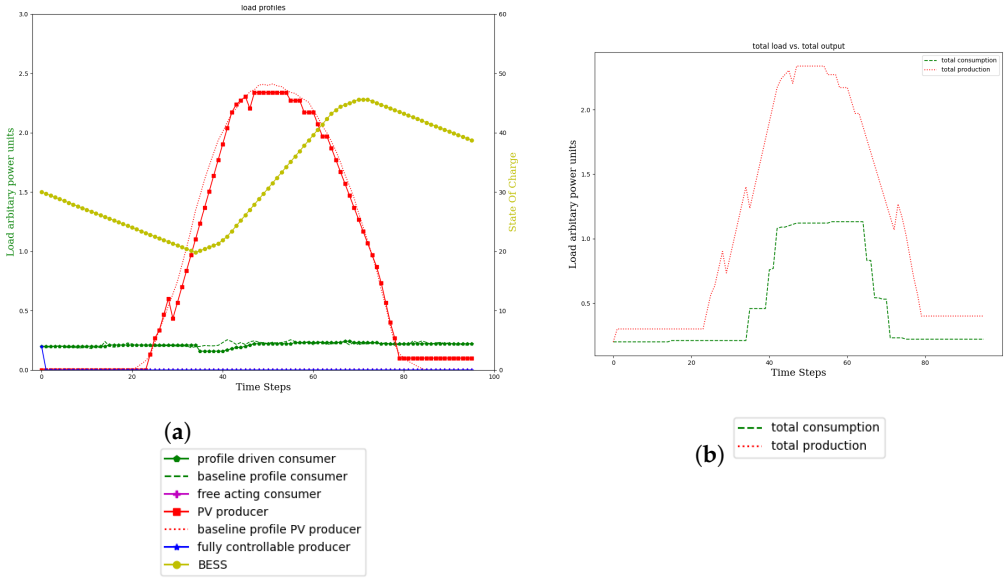


Figure 11. Evaluation on day 3, with agents trained on day 1. (**a**) Agent load curves and profiles. (**b**) Load balancing.

4.1.2. Evaluation with Theoretical Optimum

In order to provide benchmarks for the single-day trained agents, exact solutions for the problem instances day $\{1, 2, 3\}$ have been calculated with a linear program. The weights in the objective function of the linear program (cf. Equation (6)) have been set as follows: $W^{pf} = 0.5$, $W^{lb} = 0.25$, $W^{ee} = 0.25$. The evaluation results are summarised in Table 6. The numbers in the table indicate the mean absolute error between the load curve of the respective trained agent and the corresponding load curve provided by the exact solution. The range for BESS mean absolute error values is (0, 100), and for all other agents the range is (0, 1). In general, the numbers indicate good performance of the trained agents. On day 3, the FAC and the BESS agent show worse performance than on day $\{1, 2\}$. This shortcoming is mitigated with generalised training, as we will show later on in Section 4.2.

Table 6. Single-day training evaluation with benchmark solutions from the linear program.

Day	PV (mae)	Consumer (mae)	FCP (mae)	FAC (mae)	BESS (mae)
1	0.011	0.039	0.057	0.001	2.852
2	0.013	0.038	0.062	0.123	3.926
3	0.009	0.008	0.014	0.386	5.727

4.1.3. Evaluation with Stressors On

In this subsection, the focus is on resilience evaluation of the trained simulation, by confronting the trained agents with stressed energy profiles for PV producer or profile-driven consumer. All resilience experiments use problem instance day 1, cf. Table 4.

Stressing an agent means changing the agent's original profile in the time interval (40, 60), which is a period with high consumption and high renewable production, with different stressor levels low/moderate/high. For a PV profile, the stress levels are *low* (0.85), *mod* (0.48) and *high* (0.23), and the consumption profile stress levels are *low* (0.91), *mod* (0.77), *high* (0.6). The original profile's load values are multiplied with these level values, resulting in the stressed profile.

Figure 12a shows the behaviour of the trained agents with stressed PV agent. At time step 40, when the moderate stressor starts, the BESS agent starts discharging to its minimum SOC level, and the FCP agent is activated, providing positive control energy.

Since the BESS is not charged with renewable energy, the BESS agent is not able to provide enough energy in the last part of the episode, and the FCP agent is activated to achieve load balancing in the microgrid. However, during training the FCP agent was never confronted with such a stressed situation, and thus its load adaptations are not sufficient for proper load balancing. This insufficiency of the FCP agent in the last part of the episode is depicted in Figure 12b, which shows the load balancing difference curve, i.e., total production–total consumption. It points out the resilient behaviour of the trained agents: at time step 40, the stressor start is immediately followed by a distinct peak in the load balancing difference curve. However, within a few time steps the difference bounces back to within the allowed load balancing tolerance of 0.15. This is mainly due to the actions of the BESS agent, which starts discharging immediately after the stressor start. A similar behaviour can be noted at time step 60, when the stressor finishes, and the BESS starts charging.

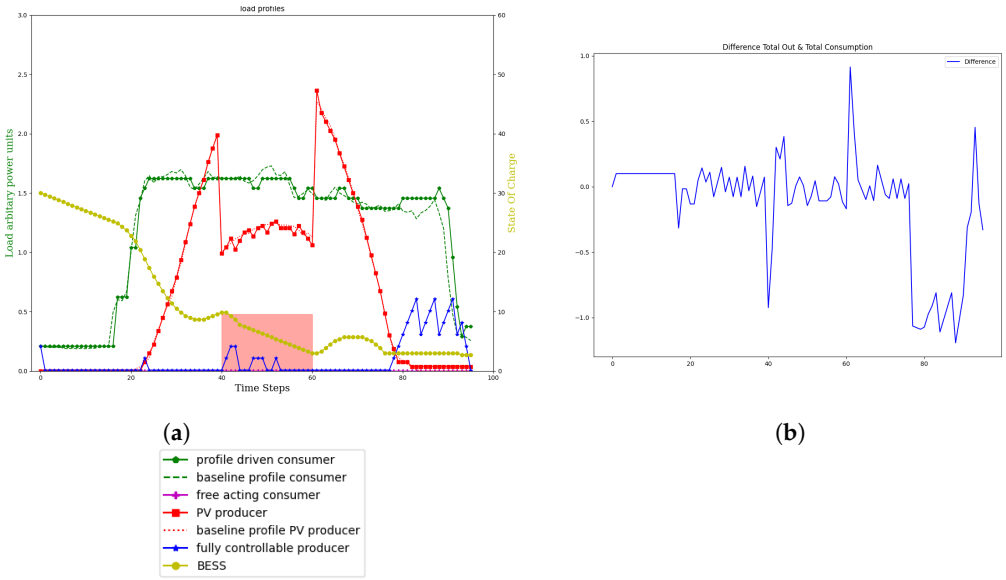


Figure 12. Evaluation with stressor on day 1, with PV agent stressed moderately. (**a**) Agent load curves and profiles. (**b**) Load balancing difference.

Figure 13a shows the behaviour of the trained agents with moderate stress on the profile-driven consumer agent between time steps (40, 60). Since the FAC agent has not been trained in cases with distinct consumption indentations, it reacts poorly and the resiliency of the microgrid is weak. This weakness can also be observed in the difference curve Figure 13b: after stressor start, load balancing never recovers between the time steps (40, 60). In Section 4.2.3 we will show that this behaviour is improved with generalised training.

Table 7 provides an overview of the quantitative analysis with the applied stressors. The table reveals the correlation between degrading FCP performance and increasing strength of the PV stressor, cf. *Diff AEE FCP* values. The values show that the FCP agent is not able to contribute sufficiently to load balancing in cases with moderate and high stress on the PV agent. Similarly, a correspondence between degrading FAC performance and increasing stress levels on consumption can be observed, cf. *Diff AEE FAC* values.

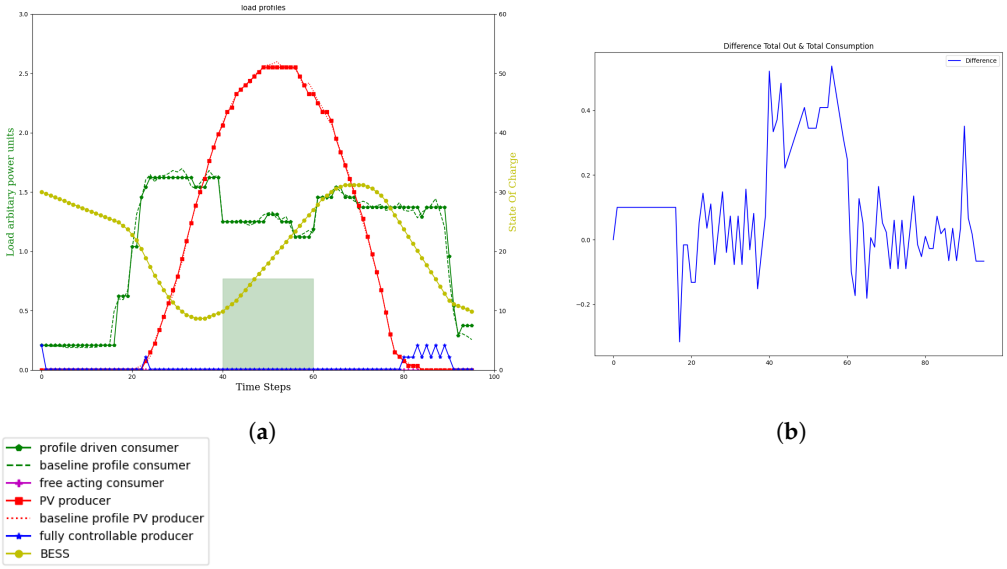


Figure 13. Evaluation of stress on day 1, with the consumer agent stressed moderately. (**a**) Agent load curves and profiles. (**b**) Load balancing difference.

Table 7. Single-day training on day 1, evaluation on day 1 with stressors.

Metric	High Stress on PV	Moderate Stress on PV	Low Stress on PV	High Stress on Consumer	Moderate Stress on Consumer	Low Stress on Consumer
Profile following PV (mae)	0.016	0.020	0.012	0.016	0.011	0.011
Profile following consumer (mae)	0.069	0.061	0.039	0.042	0.037	0.037
EE FCP	0.812	0.104	0.28	0.28	0.28	0.280
Diff AEE FCP	48.331	29.075	0.0	0.0	0.0	0.0
EE FAC	0.0	0.0	0.0	0.0	0.0	0.0
Diff AEE FAC	0.0	0.607	0.0	12.919	0.529	0.0
Storage illegal actions	0.0	0.0	0.0	0.0	0.0	0.0
Storage absolute illegal loads	0.0	0.0	0.0	0.0	0.0	0.0
Load balancing (mae)	0.361	0.260	0.089	0.212	0.146	0.097

4.1.4. Discussion Single-Day Training

The results for single-day training provide the proof-of-concept for training an EM simulation with MARL. However, the shortcomings of single-day training with respect to

generalisation are evident. A simulation trained on a specific day shows poor performance when evaluated on another day, if these days show significant differences in their energy profiles. These shortcomings will be tackled in the following section.

4.2. Generalised Training

For generalised training, the approach described in Section 3.4.4 is applied. The PV profiles all belong to the same class mentioned in Section 3.3.2, namely class 1. To compare against single-day training, evaluation is done on the same subset of three days $\{1, 2, 3\}$ from the real-world dataset, cf. Table 4. Parameters SOC_{init}, SOC_{min}, SOC_{max}, P_{max}, F_{max}, C_{max}, D_{max}, B_{max}, Δ^f, Δ^d and D_{init} remain the same as in single-day training, shown in Table 4.

4.2.1. Evaluation without Stressors

This section reports on results with respect to evaluating the generalised trained EM simulation with the three problem instances described in Table 4. The energy profiles are undisturbed, i.e., they are not modulated by stressor signals. An in-detail description of the three problem instances for evaluation can be found in Section 4.1. The evaluation results include a comparison with single-day trained simulation models.

As shown in Figures 14–16, the generalised trained agents show good EM performance, when evaluated with the different days. The agents' load curves are depicted in Figures 14a, 15a, and 16a.

Although we can observe similar good behaviour in all three days, compared to the single-day trained simulations, some slight differences appear. The main reason is the missing variance in profile data in single-day training, where the agents are prone to adapt to the given energy profiles. This can especially be seen in the comparison between the consumer profiles in Figures 7 and 14. The generalised trained consumer agent is reacting to the load changes, whereas the single-day trained agent does follow the profile much better.

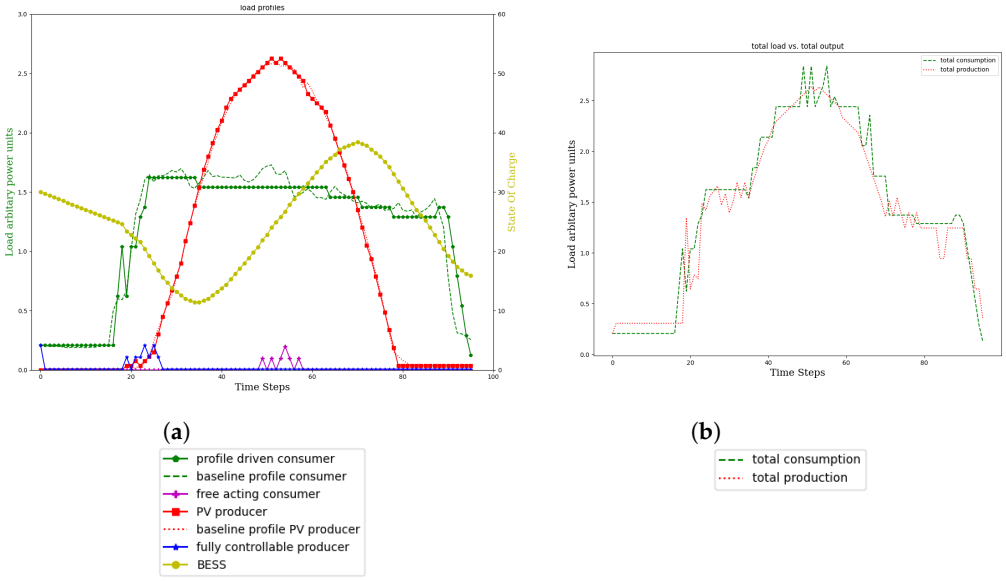


Figure 14. Evaluation on day 1, with the generalised trained agents. (**a**) Agent load curves and profiles. (**b**) Load balancing.

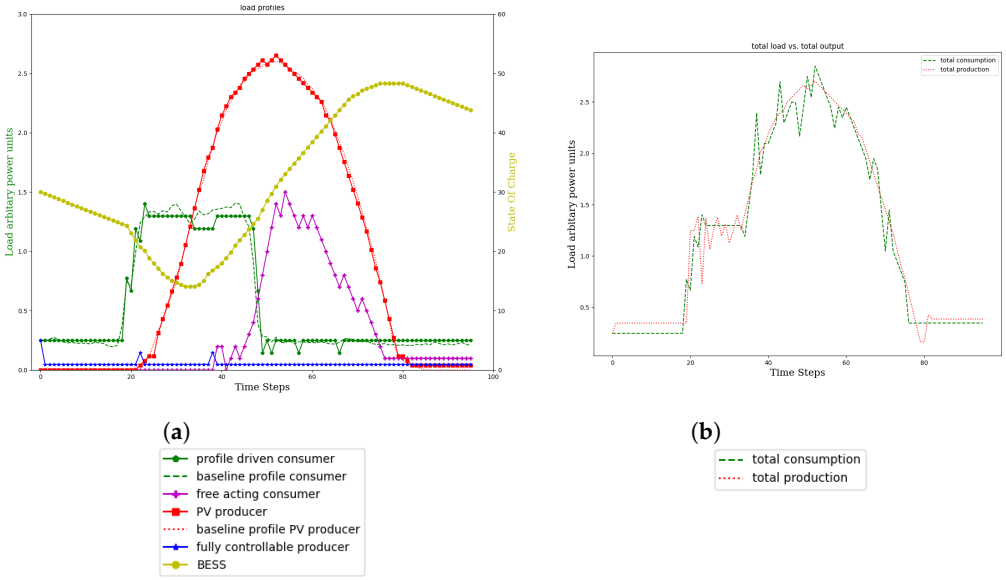


Figure 15. Evaluation on day 2, with the generalised trained agents. (**a**) Agent load curves and profiles. (**b**) Load balancing.

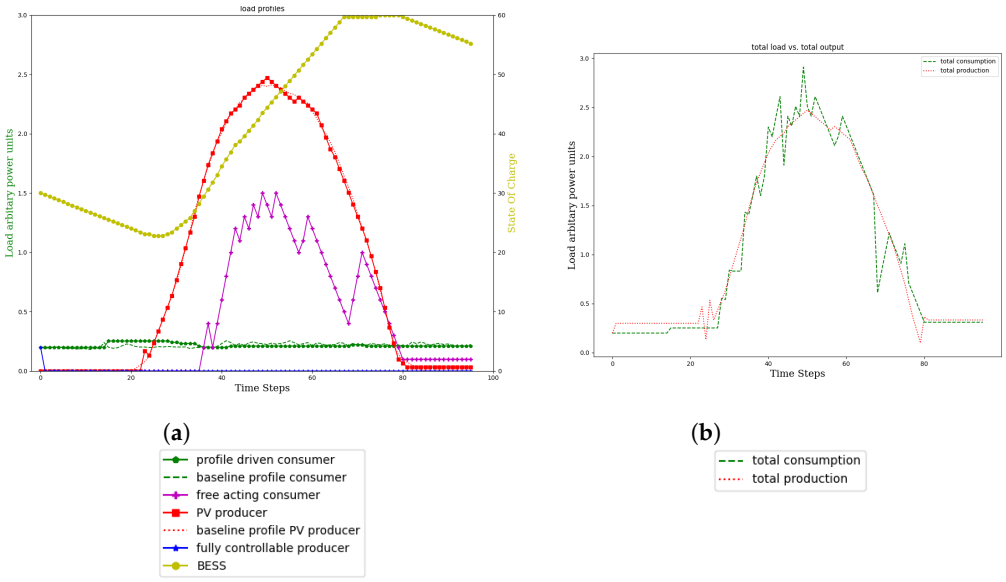


Figure 16. Evaluation on day 3, with the generalised trained agents. (**a**) Agent load curves and profiles. (**b**) Load balancing.

Through dependencies of the other agents on the profile-following agents, their behaviour is varying as well. While minor variations in the fully-controllable agents are to be expected, days 2 and 3 show more significant changes, when compared to the single-day trained simulation. On day 2, the FAC has more activity, as the BESS is not charging to its full potential. On day 3, a less fluctuating BESS leads to better load balancing possibilities for the FAC. Figures 14b, 15b and 16b illustrate that the most important EM objective, namely load balancing, is achieved well.

Keeping in mind that the results are originating from a single simulation model, they are not surprising. While it might be considered as a weakness of the generalised approach not to follow the profiles as well as in single-day training, it is more natural and leads to better results overall. More broad usage of all agents gives the needed stimulation for BESS and the fully-controllable agents to act properly. Comparing the results directly to Figures 10 and 11 shows significantly better generalisation competences, whereas the same-

day evaluated single-day trained simulation is not outperforming the general approach by far.

Table 8 shows results of the quantitative analysis of the trained agents when evaluated on the three selected days. The comparison of Tables 5 and 8 strengthens the above argument. On average, the profile-following MAE doubled, compared to single-day training and same-day evaluation. Load balancing shows better results, being on average nearly on par with the MAE of single-day trained simulations. Day 3 specifically stands out, as the generalised trained simulation shows better metrics then the single-day trained simulation.

Table 8. Generalised training quantitative evaluation for each problem instance.

Metric	Day 1 Evaluation	Day 2 Evaluation	Day 3 Evaluation
Profile-following PV (mae)	0.023	0.024	0.018
Profile-following consumer (mae)	0.084	0.05	0.021
EE FCP	0.272	2.26	0.00
Diff AEE FCP	0.0	0.596	0.0
EE FAC	0.0	1.7	1.8
Diff AEE FAC	0.0	0.715	2.367
Storage illegal actions	1.0	0.0	0.0
Storage absolute illegal loads	0.302	0.0	0.0
Load balancing (mae)	0.141	0.123	0.116

Whenever the single-day trained models are evaluated on another day than the training day, the evaluation results are significantly worse than the generalised model's results. Load balancing on average has a MAE of 3.5 times higher, which is explainable by the restricted need for fully-controllable agents in the single-day scenarios.

4.2.2. Evaluation with Theoretical Optimum

Table 9 provides an overview of quantitative analysis results with respect to comparing the generalised training results with benchmark solutions from the linear program. The benchmark solutions are the same that are used in Section 4.1.2. The comparison with Table 6 shows that the benchmark performance of the FAC agent and the BESS agent is improved with generalised training.

Table 9. Generalised training evaluation with benchmark solutions from the linear program.

Day	PV (mae)	Consumer (mae)	FCP (mae)	FAC (mae)	BESS (mae)
1	0.023	0.084	0.064	0.08	2.046
2	0.024	0.05	0.051	0.088	2.007
3	0.018	0.021	0.00	0.256	4.729

4.2.3. Evaluation with Stressors On

In this subsection, the qualitative analysis of the generalised trained agents is shown in cases with stress on profile-driven PV agent or consumer agent. The stressors are the same as in the single-day trained evaluation, cf. Section 4.1.3, and the trained simulation is evaluated on the day 1 problem instance. The stressors are active in the time interval (40, 60).

Figure 17 shows the evaluation results for the case where a moderate stressor is applied to the PV agent. The results are similar to the single-day trained agents, with resilient

behaviour after stressor start and end. The resiliency is mainly achieved by proper actions of the BESS agent. The FCP agent is active when required (i.e., in the end of the episode), but its actions are not good enough for proper load balancing. However, the *Diff AEE FCP* values in Table 10 show a significant overall improvement of the generalised trained FCP agent compared with the single-day trained FCP agent and its *Diff AEE FCP* values in Table 7.

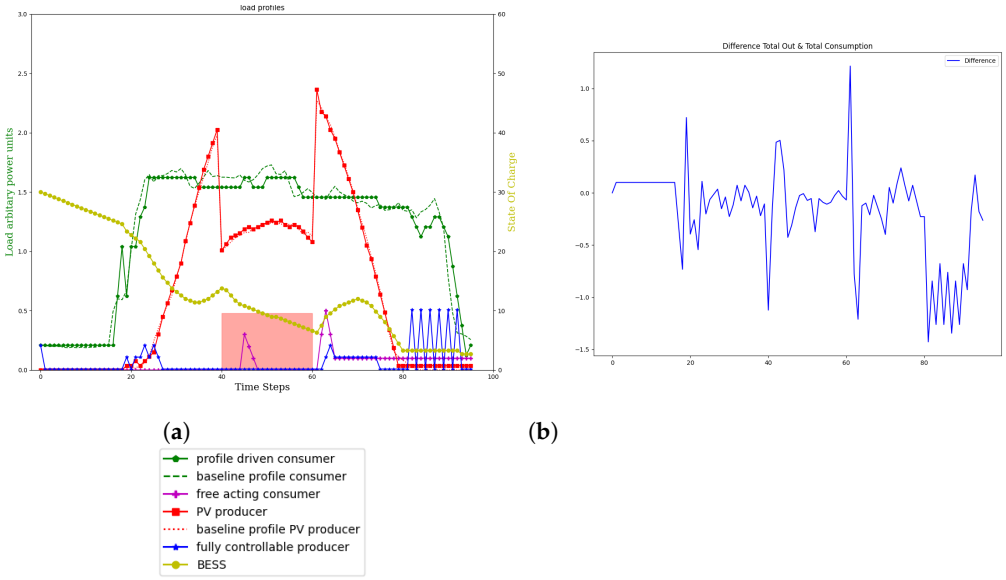


Figure 17. Evaluation of stress on day 1 (generalised), with the PV agent stressed moderately. (**a**) Agent load curves and profiles. (**b**) Load balancing difference.

In Figure 18. the agents are confronted with moderate stress on the profile-driven consumer agent. The generalised trained FAC agent acts properly during the stressed period and provides negative control energy for load balancing, which shows significant improvement compared to the single-day trained simulation, cf. Figure 13. The improvement of the FAC agent with generalised training is emphasised by the *Diff AEE FAC* values in Table 10, providing an overview of the quantitative evaluation results for various stressors.

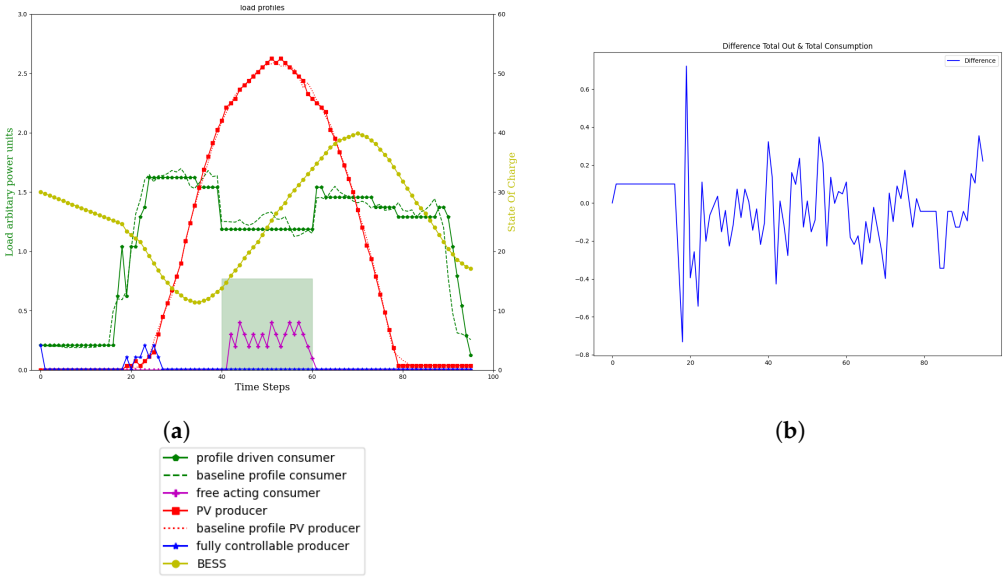


Figure 18. Evaluation of stress on day 1 (generalised), with consumer agent stressed moderately. (**a**) Agent load curves and profiles. (**b**) Load balancing difference.

Table 10. Generalised training, evaluation on day 1 with the stressors switched on.

Metric	High Stress on PV	Moderate Stress on PV	Low Stress on PV	High Stress on Consumer	Moderate Stress on Consumer	Low Stress on Consumer
Profile following PV (mae)	0.020	0.023	0.024	0.024	0.024	0.025
Profile following consumer (mae)	0.084	0.074	0.084	0.074	0.079	0.080
EE FCP	2.112	0.904	0.264	0.272	0.272	0.272
Diff AEE FCP	29.422	16.706	0.0	0.0	0.0	0.0
EE FAC	0.0	3.20	0.0	0.0	0.0	0.0
Diff AEE FAC	3.354	1.081	0.0	0.0	0.0	0.0
Storage illegal actions	1.0	1.0	2.0	1.0	1.0	1.0
Storage absolute illegal loads	0.302	0.302	0.641	0.302	0.302	0.302
Load balancing (mae)	0.356	0.296	0.158	0.158	0.150	0.144

4.2.4. Discussion Generalised Training

The results for generalised training show that it is possible to train a single simulation model that shows good EM performance for the complete set of problem instances $\{1, 2, 3\}$. With the generalised training approach, the generalisation capabilities of the agents are significantly improved, compared to single-day trained agents.

Tables 11and 12 show that improved generalisation does also result in better reliability and resilience. In Table 12 four simulation models are compared: three single-day trained models, and one generalised trained model. The inverse reliability scores are calculated according to Equation (7), with days $\{1, 2, 3\}$ as the set of problem instances. The reliability advantage of the generalised trained model with score 0.127 is evident.

Table 11 provides an overview of inverse resilience scores for various stressors. Two simulation models are compared: (1) agents trained on day 1, and (2) generalised trained agents. Both models were not confronted with stressors during training. We have investigated the impact of low/moderate/high stressors on PV producer and profile-driven consumer, respectively. The inverse resilience scores are calculated according to Equation (8). The scores show that the generalised trained agents outperform the single day trained agents for each of the stressors.

Table 11. Inverse resilience scores for single day training and generalised training.

Stressor	Day 1 Trained Inverse Resilience Score	Generalised Trained Inverse Resilience Score
High stress on PV	3.56	1.52
Moderate stress on PV	2.29	1.09
Low stress on PV	0.126	0.12
High stress on consumer	1.68	0.12
Moderate stress on consumer	4.26	0.06
Low stress on consumer	0.227	0.021

Table 12. Inverse reliability scores for single day training and generalised training

Trained Model	Inverse Reliability Score
Day 1 trained	0.387
Day 2 trained	0.318
Day 3 trained	0.306
Generalised trained	0.127

5. Conclusions

Our contribution allows for the training of an EM simulation for microgrids with MARL, leveraging real-world energy profile data for energy consumption and renewable energy production. The quantitative analysis, including the comparison with exact solutions for the EM problem from a linear program, shows good performance of the trained simulation. It will help decision makers to estimate the resilience and reliability of the microgrid system.

In more detail, the contributions of this work can be summarized as follows:

- A data-driven EM simulation has been developed, which is able to demonstrate reliable and resilient behaviour;
- To make the simulation adapt to different data patterns, a strong emphasis on generalisation capabilities have been given. This included, analysis of real-world energy profiles, generated artificial energy profiles and generalised training with artificial profiles;
- A multi-dimensional reward scheme was developed to incorporate different EM performances such as load balancing, maximise usage of renewable energy, minimise usage of non-renewable energy, battery charging/discharging behaviour, and energy profile-following;
- An extension of the actor-critic RL algorithm PPO was employed with centralised critic to deal with the collaborative behaviour of each agent;
- As a result of strong generalisation capabilities, the realisation of reliability was noted;

- Finally, evaluation of trained agents with stressors switched on provided insights on the resilience aspects of trained agents in a RE microgrid.

An important development target was "trained agents with good generalisation capabilities". This means that the trained EM simulation should exhibit good performance for a multitude of EM problem instances. The presented results are based on a set of three problem instances, with considerable differences in the energy consumption profiles. In further research, the number of problem instances and the variety of data patterns in energy profiles will be increased, thus continuing the work towards the above development target.

Aiming at good generalisation capabilities, a specific methodology was developed. The methodology comprises analysis of real-world energy profile data, generation of artificial profile data based on the analysis, and training with the artificial data. Computational results demonstrate the impact of this training methodology on EM performance, especially with respect to reliability and resilience. An EM simulation with good generalisation capability shows better reliability, and better resilience in the face of stressors, even if these stressors have never been encountered in the training phase of the EM simulation.

In RL, design and parametrisation of the reward scheme are crucial development steps. A lesson learned is the need for iterative cycles of reward scheme design, tuning the reward parameters with hyperparameter search, and evaluation. For a complex MARL task with a multi-dimensional training objective (such as EM in microgrids) it is not sufficient to apply a methodology where huge efforts are put into the design phase, and after the design phase the reward parameters are tuned in iterative hyperparameter search and evaluation cycles.

In this contribution, the focus is on training an EM simulation, where training and execution of the simulation are done in a centralised fashion. This means that the agents that make up the simulation are trained/executed in a central training/execution environment. However, the aim of future research is to build on the results described in this work, and to develop an EM system that actually manages the loads in a physical microgrid. We envisage the application of the "centralised training and decentralised deployment" paradigm: the agents are trained in a central training environment, and after training they are deployed on the microgrid in a decentralised fashion. For example, a BESS agent would be deployed on the local control system for the corresponding BESS component in the microgrid.

In the presented work, a microgrid consisting of five components is used. In real-world EM deployment use cases, the number of microgrid components, and thus the number of agents to be trained, could be significantly larger than five. From the point of view of MARL, a research question is then: "How does the EM training performance in MARL scale with the number of agents to be trained?" Another research challenge that comes in the wake of the deployment scenario is the integration of the trained agents with the cyber-physical microgrid infrastructure. In this work, the agents' observations from the environment, and the effects of agents' actions on the environment, are simulated. In a real microgrid, the observations for a deployed agent must be provided by the cyber-physical infrastructure, sensing and transmitting physical parameter values to the agent. Similarly, the deployed agent's actions have to be translated into actual load adaptation signals for the corresponding microgrid component.

Author Contributions: Conceptualization, K.D., A.H. and P.M.; methodology, K.D., A.H. and P.M.; software, K.D., A.H., P.M. and H.Z.; validation, K.D., A.H. and P.M.; formal analysis, K.D., A.H., P.M. and H.Z.; investigation, K.D., A.H., P.M. and H.Z.; resources, A.H., H.Z. and G.W.; data curation, P.M., K.D.; writing-original draft preparation, K.D., A.H., P.M. and H.Z.; writing—review and editing, A.H.,G.W.; visualisation, K.D., A.H., P.M. and H.Z.; supervision, A.P.; principal investigator, G.W.; funding acquisition, A.P. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research has been partially funded by the European Commission and the Government of Upper Austria through the program EFRE/IWB 2020, the priority REACT-EU, and the project RESINET (RESIlienzsteigerung In energieNETzen—RESilience increase In energy NETworks; Nr.:WI-2020-701900/12). It has also been supported by *Pro*^{2}*Future* (contract No. 854184). *Pro*^{2}*Future* is

funded within the Austrian COMET Program—Competence Centers for Excellent Technologies—under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry for Digital and Economic Affairs and of the Provinces of Upper Austria and Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

Data Availability Statement: Data for this study were provided by the University of Applied Science Upper Austria, Wels campus.

Acknowledgments: The authors thank the colleagues from the University of Applied Science Upper Austria, RISC, RECENDT for the fruitful discussions. We also thank the colleagues for making the data available.

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

Appendix A. Tool to Extract Energy Profiles

In the project, we used datasets of photovoltaic systems from Upper Austria. One dataset contains data from about 190 private smaller photovoltaic systems equipped with battery packs. In addition, a second dataset from larger systems *'Grossanlagen'* was also provided. The records are from the period 2014 to 2021, with a recording interval of 15 or 5 min, and were provided as MATLAB export files. In total, these datasets contain over 100 million rows of data. Constant parameters available for each system are zip code, model, and capacity. The available properties are PV production, power consumption, grid power, grid injection, how much battery charged, how much battery discharged, and battery SOC .

To test our algorithms, we wanted to extract interesting daily profiles for electricity generation and consumption from this data. For this task, a profile extraction tool was developed, that first allows to get an overview of the existing data. The data for the small pv systems are available as matlab export files. However, the *'Grossanlagen'* dataset uses the HDF5 file format, which is used in MATLAB export files > version 3.7. Therefore, the application is able to read both formats of data.

To quickly develop the application, we decided to use python. With the help of Plotly-Dash (https://dash.plotly.com/, accessed on 2 February 2022) we created some appropriate dashboards. Users in this application can first select whether they want to examine the data of large or small PV systems. The next step is to select the concrete system. The charts now show an overview of the data of the PV and the consumers in the entire time range (see Figure A1). In addition, two drop down menus are filled with all months and days for which data are included in the time series. The drop down menus can be used to filter data accordingly.

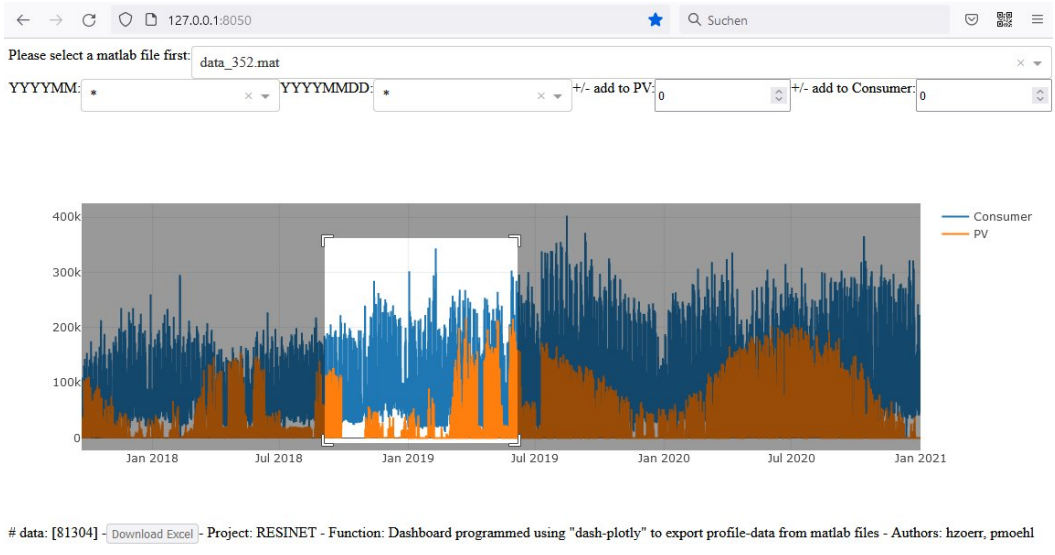


Figure A1. Overview profile extraction tool.

Users can filter individual months (e.g., Figure A2) and search the data for days of interest (e.g., Figure A3).

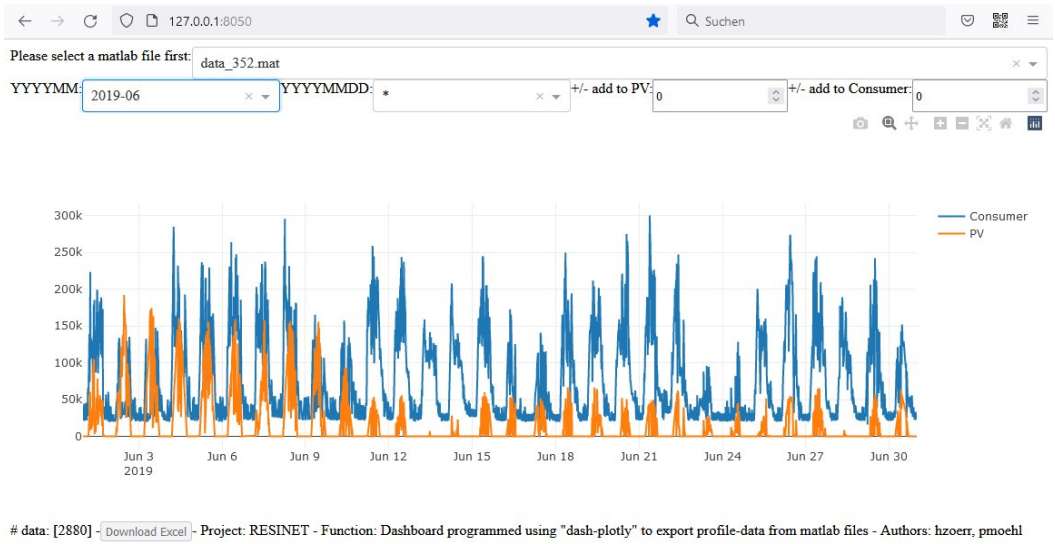


Figure A2. A sample month in the profile extraction tool.

If exactly one day is selected as a filter, export of the data is then possible. Users can then activate the *Download Excel* button and the day's data will be exported to MS-Excel as a daily profile, and can be used in experiments.

The final profile contains values for every 15 min of the day (so the dataset contains exactly 96 data points).

This extraction tool has proven to be very useful in the initial stages of the project.

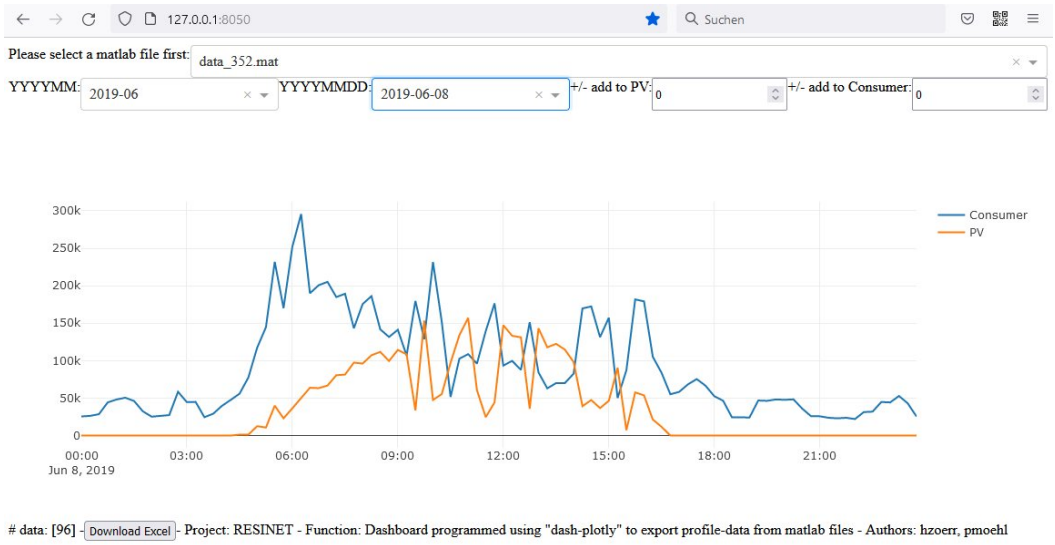


Figure A3. A sample day in the profile extraction tool.

References

1. Liu, C.; Li, D.; Wang, L.; Li, L.; Wang, K. Strong robustness and high accuracy in predicting remaining useful life of supercapacitors. *APL Mater.* **2022**, *10*, 061106. [CrossRef]
2. Cui, Z.; Kang, L.; Li, L.; Wang, L.; Wang, K. A combined state-of-charge estimation method for lithium-ion battery using an improved BGRU network and UKF. *Energy* **2022**, *259*, 124933. [CrossRef]
3. Holappa, L. A General Vision for Reduction of Energy Consumption and CO_2 Emissions from the Steel Industry. *Metals* **2020**, *10*, 1117. [CrossRef]
4. Fronius International GmbH. Microgrid with Fronius Inverter at a Fish Farm in Mali. 2020. Available online: https://www.fronius.com/en/solar-energy/installers-partners/infocentre/references/mli-djoliba (accessed on 28 July 2022).

5. Hillig, T. Rural Electrification in Times of Corona. 2020. Available online: https://www.pv-tech.org/rural-electrification-in-times-of-corona/ (accessed on 28 July 2022).
6. Ellis, M. Smart Grid: The Components and Integrating Communication. In Proceedings of the 2012 IEEE Green Technologies Conference, Tulsa, OK, USA, 19–20 April 2012; pp. 1–6. [CrossRef]
7. Panteli, M.; Mancarella, P.; Trakas, D.N.; Kyriakides, E.; Hatziargyriou, N.D. Metrics and Quantification of Operational and Infrastructure Resilience in Power Systems. *IEEE Trans. Power Syst.* **2017**, *32*, 4732–4742. [CrossRef]
8. Zhang, D.; Han, X.; Deng, C. Review on the Research and Practice of Deep Learning and Reinforcement Learning in Smart Grids. *CSEE J. Power Energy Syst.* **2018**, *4*, 362–370. [CrossRef]
9. IEEE. *IEEE Standard for the Specification of Microgrid Controllers*; IEEE: Piscataway, NJ, USA, 2017.
10. Hussain, A.; Bui, V.H.; Kim, H.M. Microgrids as a resilience resource and strategies used by microgrids for enhancing resilience. *Appl. Energy* **2019**, *240*, 56–72. [CrossRef]
11. Strbac, G.; Hatziargyriou, N.; Lopes, J.P.; Moreira, C.; Dimeas, A.; Papadaskalopoulos, D. Microgrids: Enhancing the Resilience of the European Megagrid. *IEEE Power Energy Mag.* **2015**, *13*, 35–43. [CrossRef]
12. Abu-Elzait, S.; Parkin, R. Economic and environmental advantages of renewable-based microgrids over conventional microgrids. In Proceedings of the 2019 IEEE Green Technologies Conference (GreenTech), Lafayette, LA, USA, 3–6 April 2019; pp. 1–4.
13. Chaudhary, G.; Lamb, J.J.; Burheim, O.S.; Austbø, B. Review of Energy Storage and Energy Management System Control Strategies in Microgrids. *Energies* **2021**, *14*, 4929. [CrossRef]
14. Perera, A.; Nik, V.M.; Chen, D.; Scartezzini, J.L.; Hong, T. Quantifying the impacts of climate change and extreme climate events on energy systems. *Nat. Energy* **2020**, *5*, 150–159. [CrossRef]
15. Jia, Q.S.; Panetto, H.; Macchi, M.; Siri, S.; Weichhart, G.; Xu, Z. Control for Smart Systems: Challenges and Trends in Smart Cities. *Annu. Rev. Control* **2022**, *53*, 358–369. [CrossRef]
16. Bajwa, A.A.; Mokhlis, H.; Mekhilef, S.; Mubin, M. Enhancing power system resilience leveraging microgrids: A review. *J. Renew. Sustain. Energy* **2019**, *11*, 035503. [CrossRef]
17. Jasiūnas, J.; Lund, P.D.; Mikkola, J. Energy system resilience—A review. *Renew. Sustain. Energy Rev.* **2021**, *150*, 111476. [CrossRef]
18. Weichhart, G. Enterprise Integration and Interoperability improving Business Analytics. In Proceedings of the 2nd International Conference on Innovative Intelligent Industrial Production and Logistics, Insticc, online, 25–27 October 2021; pp. 227–236.
19. Abbey, C.; Cornforth, D.; Hatziargyriou, N.; Hirose, K.; Kwasinski, A.; Kyriakides, E.; Platt, G.; Reyes, L.; Suryanarayanan, S. Powering Through the Storm: Microgrids Operation for More Efficient Disaster Recovery. *IEEE Power Energy Mag.* **2014**, *12*, 67–76. [CrossRef]
20. Clark-Ginsberg, A. *What's the Difference between Reliability and Resilience*; Department of Homeland Security, Stanford University: Stanford, CA, USA, 2016.
21. Cuadra, L.; Salcedo-Sanz, S.; Del Ser, J.; Jiménez-Fernández, S.; Geem, Z.W. A critical review of robustness in power grids using complex networks concepts. *Energies* **2015**, *8*, 9211–9265. [CrossRef]
22. Amani, A.M.; Jalili, M. Power Grids as Complex Networks: Resilience and Reliability Analysis. *IEEE Access* **2021**, *9*, 119010–119031. [CrossRef]
23. Wang, Y.F. Adaptive job shop scheduling strategy based on weighted Q-learning algorithm. *J. Intell. Manuf.* **2020**, *31*, 417–432. [CrossRef]
24. Mujjuni, F.; Betts, T.; To, L.; Blanchard, R. Resilience a means to development: A resilience assessment framework and a catalogue of indicators. *Renew. Sustain. Energy Rev.* **2021**, *152*, 111684. [CrossRef]
25. Panteli, M.; Pickering, C.; Wilkinson, S.; Dawson, R.; Mancarella, P. Power System Resilience to Extreme Weather: Fragility Modeling, Probabilistic Impact Assessment, and Adaptation Measures. *IEEE Trans. Power Syst.* **2017**, *32*, 3747–3757. [CrossRef]
26. Dehghani, N.L.; Jeddi, A.B.; Shafieezadeh, A. Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning. *Appl. Energy* **2021**, *285*, 116355. [CrossRef]
27. Jufri, F.H.; Widiputra, V.; Jung, J. State-of-the-art review on power grid resilience to extreme weather events: Definitions, frameworks, quantitative assessment methodologies, and enhancement strategies. *Appl. Energy* **2019**, *239*, 1049–1065. [CrossRef]
28. Huang, G.; Wang, J.; Chen, C.; Guo, C.; Zhu, B. System resilience enhancement: Smart grid and beyond. *Front. Eng. Manag.* **2017**, *4*, 271. [CrossRef]
29. Wei, F.; Wan, Z.; He, H. Cyber-Attack Recovery Strategy for Smart Grid Based on Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2020**, *11*, 2476–2486. [CrossRef]
30. Weichhart, G.; Mangler, J.; Raschendorfer, A.; Mayr-Dorn, C.; Huemer, C.; Hämmerle, A.; Pichler, A. An Adaptive System-of-Systems Approach for Resilient Manufacturing. *e & i Elektrotechnik Informationstechnik* **2021**, *138*, 341–348. [CrossRef]
31. Muriithi, G.; Chowdhury, S. Optimal Energy Management of a Grid-Tied Solar PV-Battery Microgrid: A Reinforcement Learning Approach. *Energies* **2021**, *14*, 2700. [CrossRef]
32. Zhang, Z.; Zhang, D.; Qiu, R. Deep reinforcement learning for power system: An overview. *CSEE J. Power Energy Syst.* **2019**, *6*, 213–225. [CrossRef]
33. Lee, S.; Xie, L.; Choi, D.H. Privacy-Preserving Energy Management of a Shared Energy Storage System for Smart Buildings: A Federated Deep Reinforcement Learning Approach. *Sensors* **2021**, *21*, 4898. [CrossRef]

34. Hämmerle, A.; Deshpande, K.; Möhl, P.; Weichhart, G. Training an Energy Management Simulation using Multi-Agent Reinforcement Learning. In Proceedings of the ENERGY 2022—The Twelfth International Conference on Smart Grids, Green Communications and IT Energy-Aware Technologies, Venice, Italy, 22–26 May 2022.
35. Qin, Z.; Liu, D.; Hua, H.; Cao, J. Privacy Preserving Load Control of Residential Microgrid via Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2021**, *12*, 4079–4089. [CrossRef]
36. Ji, Y.; Wang, J.; Xu, J.; Fang, X.; Zhang, H. Real-Time Energy Management of a Microgrid Using Deep Reinforcement Learning. *Energies* **2019**, *12*, 2291. [CrossRef]
37. Ali, K.H.; Sigalo, M.; Das, S.; Anderlini, E.; Tahir, A.A.; Abusara, M. Reinforcement Learning for Energy-Storage Systems in Grid-Connected Microgrids: An Investigation of Online vs. Offline Implementation. *Energies* **2021**, *14*, 5688. [CrossRef]
38. Samadi, E.; Badri, A.; Ebrahimpour, R. Decentralized multi-agent based energy management of microgrid using reinforcement learning. *Int. J. Electr. Power Energy Syst.* **2020**, *122*, 106211. [CrossRef]
39. Foruzan, E.; Soh, L.K.; Asgarpoor, S. Reinforcement Learning Approach for Optimal Distributed Energy Management in a Microgrid. *IEEE Trans. Power Syst.* **2018**, *33*, 5749–5758. [CrossRef]
40. Fang, X.; Wang, J.; Song, G.; Han, Y.; Zhao, Q.; Cao, Z. Multi-Agent Reinforcement Learning Approach for Residential Microgrid Energy Scheduling. *Energies* **2019**, *13*, 123. [CrossRef]
41. Fang, X.; Zhao, Q.; Wang, J.; Han, Y.; Li, Y. Multi-agent Deep Reinforcement Learning for Distributed Energy Management and Strategy Optimization of Microgrid Market. *Sustain. Cities Soc.* **2021**, *74*, 103163. [CrossRef]
42. Panfili, M.; Giuseppi, A.; Fiaschetti, A.; Al-Jibreen, H.B.; Pietrabissa, A.; Delli Priscoli, F. A Game-Theoretical Approach to Cyber-Security of Critical Infrastructures Based on Multi-Agent Reinforcement Learning. In Proceedings of the 2018 26th Mediterranean Conference on Control and Automation (MED), Zadar, Croatia, 19–22 June 2018; pp. 460–465. [CrossRef]
43. Qazi, H.S.; Liu, N.; Wang, T. Coordinated energy and reserve sharing of isolated microgrid cluster using deep reinforcement learning. In Proceedings of the 2018 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE), Chengdu, China, 4–7 June 2020; pp. 81–86.
44. Zhao, J.; Li, F.; Mukherjee, S.; Sticht, C. Deep Reinforcement Learning based Model-free On-line Dynamic Multi-Microgrid Formation to Enhance Resilience. *IEEE Trans. Smart Grid* **2022**, *13*, 2557–2567. [CrossRef]
45. Nie, H.; Chen, Y.; Xia, Y.; Huang, S.; Liu, B. Optimizing the Post-Disaster Control of Islanded Microgrid: A Multi-Agent Deep Reinforcement Learning Approach. *IEEE Access* **2020**, *8*, 153455–153469. [CrossRef]
46. Kamruzzaman, M.; Duan, J.; Shi, D.; Benidris, M. A Deep Reinforcement Learning-Based Multi-Agent Framework to Enhance Power System Resilience Using Shunt Resources. *IEEE Trans. Power Syst.* **2021**, *36*, 5525–5536. [CrossRef]
47. Schulman, J.; Klimov, O.; Wolski, F.; Dhariwal, P.; Radford, A. Proximal Policy Optimization. 2017. Available online: https://openai.com/blog/openai-baselines-ppo/ (accessed on 29 July 2022).
48. Zdravković, M.; Panetto, H.; Weichhart, G. AI-enabled Enterprise Information Systems for Manufacturing. *Enterp. Inf. Syst.* **2021**, *16*, 668–720. [CrossRef]
49. Thalmann, S.; Mangler, J.; Schreck, T.; Huemer, C.; Streit, M.; Pauker, F.; Weichhart, G.; Schulte, S.; Kittl, C.; Pollak, C.; et al. Data Analytics for Industrial Process Improvement A Vision Paper. In Proceedings of the 2018 IEEE 20th Conference on Business Informatics (CBI), Vienna, Austria, 11–14 July 2018; IEEE Computer Society: Los Alamitos, CA, USA, 2018; Volume 1, pp. 92–96. [CrossRef]
50. Weichhart, G.; Panetto, H.; Gutiérrez, A.M. Interoperability in the Cyber-Physical Manufacturing Enterprise. *Annu. Rev. Control* **2021**, *51*, 346–356. [CrossRef]
51. Baird, L.; Moore, A. Gradient Descent for General Reinforcement Learning. In *Advances in Neura Information Processing Systems*; Kearns, M., Solla, S., Cohn, D., Eds.; MIT Press: Cambridge, MA, USA, 1998; Volume 11.
52. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
53. Yu, C.; Velu, A.; Vinitsky, E.; Wang, Y.; Bayen, A.; Wu, Y. The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. *arXiv* **2021**, arXiv:2103.01955.
54. 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, Stockholm Sweden, 10–15 July 2018; Volume 80, pp. 3053–3062.
55. Tang, Y.; Agrawal, S. Discretizing Continuous Action Space for On-Policy Optimization. *arXiv* **2020**, arXiv:1901.10500.