IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 March 2023, accepted 4 April 2023, date of publication 17 April 2023, date of current version 26 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3267978

RESEARCH ARTICLE

Minimizing Energy Cost in PV Battery Storage Systems Using Reinforcement Learning

FLORUS HÄRTEL AND THILO BOCKLISCH
Professur für Energiespeichersysteme, Technische Universität Dresden, 01069 Dresden, Germany

Corresponding author: Florus Härtel (florus.haertel@tu-dresden.de)

This work was supported in part by the Research Project "HYBAT – Hybrid Lithium-Ion Battery Storage Solution with 1500 V System Technology, Innovative Thermal Management and Optimized System Management", and in part by the Federal Ministry of Economic Affairs and Climate Action under Grant 03EI3009C.

ABSTRACT This article addresses the development and tuning of an energy management for a photovoltaic (PV) battery storage system for the cost-optimized use of PV energy using reinforcement learning (RL). An innovative energy management concept based on the Proximal Policy Optimization algorithm in combination with recurrent Long Short-Term Memory neural networks is developed for data-based policy learning, a concept that has been rarely addressed in the literature so far. As a reference system for the simulation-based investigations, a PV battery storage system is modelled, parametrized and implemented with an interface for the RL algorithm. To demonstrate the generalization capability of the learned energy management, 98 training and 12 evaluation episodes, each with a length of one year, are generated from an empirical dataset of global radiation and load power time series. To improve the convergence speed and stability of the RL algorithm as well as the learned policy with regards to techno-economic metrics, an extensive hyperparameter study is conducted by training 216 control policies with different hyperparameter configurations. A simulation-based evaluation of the learned energy management against conventional rule-based and model-predictive energy managements shows that the RL-based concept can achieve slightly better results in terms of energy costs and the amount of energy fed into the grid than the commonly used model-predictive method.

INDEX TERMS PV battery storage system (PVBSS), energy storage system (ESS), reinforcement learning (RL), energy management (EM), proximal policy optimization (PPO), long short-term memory (LSTM), optimal control, hyperparameter tuning.

I. INTRODUCTION

The increasing share of photovoltaic and wind power in the energy mix of Germany due to the legally defined climate protection targets of the federal government [1] is leading to an increasing demand for stationary battery storage systems [2] to compensate for the volatility of renewable electricity sources and thereby ensuring the stability of the grid [2]. The reference expansion scenario of the Fraunhofer Institute for Solar Energy Systems ISE predicts 104 GWh of installed stationary battery storage capacity by 2030 in Germany for this purpose [3]. Thereof, only 5.43 GWh have been installed by December 2022 [4]. In addition to this enormous expansion demand, falling manufacturing costs of battery storage technologies over the last decade have made their deployment increasingly attractive from an economic point of view [5], [6], [7]. In order to fulfill the techno-economic objectives for the deployment of a battery storage system, such as low operating costs or a short amortization period, intelligent energy management (EM) concepts are required that can take into account several opposing optimization objectives simultaneously, for example, maximizing self-sufficiency and maximizing energy fed into the grid at the same time.

Reinforcement learning (RL) is a model-free method that can be used to optimize a control policy, in this case the EM of the PV battery storage system (PVBSS), by interacting with this system and thereby receiving and optimizing a reward [8], [9]. Generally, RL offers a number of advantages over the widely used model-predictive control (MPC)

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Wang.

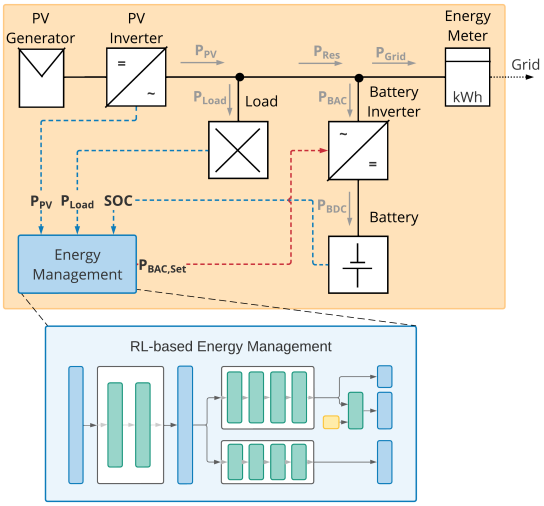

FIGURE 1. AC-coupled PVBSS with input values (blue) and output values (red) of the energy management modelled for the simulation-based investigations. The energy management can be exchanged modularly in the simulation-based investigations. The neural network of the RL-based energy management is shown in simplified representation as one example.

methods: A) In contrast to MPC methods, no explicit system and forecast models need to be formulated. They can be thought of as being part of the policy itself and are therefore learned by the RL algorithm. B) The policy can be optimized effectively over an infinite time horizon via discounted rewards and a state value function, which is also learned by RL algorithm. C) No restrictions need to be imposed on the modelling of the environment, i.e. the PVBSS, or the reward function as they are regarded as black-box functions. D) Compared to MPC methods, less computational power is needed during online operation, since the policy optimization is carried out prior to the application and no real-time optimization is needed. RL has been applied successfully to a variety of control problems in robotics, navigation and power systems [10], [11], which would have been very hard to solve with MPC, due to the complexity of modelling the system and solving the corresponding optimization problem in real-time.

In this paper, an RL-based EM concept using the Proximal Policy Optimization (PPO) algorithm in combination with Long Short-Term Memory (LSTM) networks for the cost-optimal power allocation in a PVBSS (see Figure 1) is developed. Its configuration and tuning is highlighted and its performance is compared with a conventional priority-based and a model-predictive EM concept. In section II, the state of research and technology is considered and the control problem for the EM is formulated. Section III describes the PVBSS with a PV system, a cumulative load, a battery storage, a grid connection and the EM. As interface for the RL algorithm the state, action and reward are defined. Section IV details the developed EM concept based on the PPO algorithm. An extensive simulation-based study on the hyperparameter settings of the PPO algorithm is conducted using the High Performance Computing cluster of TU Dresden. Section V evaluates the learned control policy against a rule-based and a model-predictive EM concept. Section VI summarizes the results of the investigations and provides an outlook on further research questions in this area.

II. STATE OF THE ART AND PROBLEM DEFINITION

A. ENERGY MANAGEMENT FOR PV BATTERY STORAGE SYSTEMS

Typical techno-economic objectives for the operation of a PVBSS are to maximize the degree of self-sufficiency, to minimize PV curtailment, to optimize the lifetime of the battery storage or to maximize the overall system efficiency [12]. The task of the EM is to allocate the power flow in a PVBSS optimally with regard to one or many of these criteria. A large number of rule-based and model-predictive EM concepts can be found in the literature to address this task [13]. Generally, an implemented EM receives a measurement of the relevant system state variables as input and outputs the setpoint values for an underlying control circuit, i.e. the controller of the battery inverter, operating at a much higher frequency than the EM itself.

Rule-based EM concepts calculate theses setpoint values, i.e. the battery power setpoint or grid power setpoint, according to simple logical or arithmetic operations, for example by calculating the residual load or by triggering actions when crossing a defined threshold of the battery state of charge (SOC). These pre-defined rules are derived analytically ahead of online operation. Forecasts and real-time optimization are therefore not regarded. The advantage of this EM concept is the low complexity of implementation and the low computing requirements in online operation. The simple and commonly used priority-based EM usually provides the best results with regard to the degree of self-sufficiency of a PVBSS [12], [14]. However, rule-based EM concepts cannot be used for more complex EM tasks, such as the multi-objective optimization [15] discussed in this paper, or the management of hybrid energy storage systems [16], since an analytical derivation of an optimal control policy is not practical for these cases.

Model-predictive EM concepts optimize a trajectory of setpoint values with regard to a defined objective function over a time window projecting into the future. PV and load forecasts as well as a model of the PVBSS need to be formulated, in order to predict the system behavior. The forecasts and optimization are periodically updated in order to correct prediction errors.

A range of variations of the model-predictive EM concept [13] with different optimization methods, such as linear programming [17] or dynamic programming [12], [18], and different forecasting methods [19] have been investigated in the literature. Model-predictive EM concepts are widely regarded as state of the art in this field. However, the explicit modeling of the PVBSS and the forecast of the PV and load power are challenges in the application of these methods. Both increase the implementation complexity and represents a systematic source of error due to deviations between the forecast or optimization model and the actual system. Moreover, the real-time optimization required for this method scales unfavorably for large and non-linear control problems.

In this paper the priority-based EM concept (PRIO) serves as a reference for evaluating the performance of the developed RL-based EM concept as it is the simplest form of an EM for a PVBSS and still widely used in real-world applications. Further, a model-predictive energy management (MPC) [20] is considered as second reference EM, because of its explicit maximization of the energy fed into the grid and minimization of the energy cost.

B. REINFORCEMENT LEARNING IN ENERGY STORAGE APPLICATIONS

A third EM concept is based on the model-free RL method, offering a range of advantages over MPC methods. The EM of the PVBSS is regarded as a Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}_a, \mathcal{R}_a, \gamma)$. Where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space and $\mathcal{P}_a(s_{t+1} \mid s_t, a_t) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the stochastic state transition function from state s_t to s_{t+1} when action a_t is executed. $\mathcal{R}_a(s, a) : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ is the reward function and γ is the discount factor.

In order to solve the MDP with RL, first, a trajectory $\tau = (s_0, a_0, r_0, \ldots a_{T-1}, r_{T-1}, s_T)$ is generated by executing the current policy $a_t \sim \pi_\theta(\cdot|s_t, \theta)$ of the RL agent in interaction with the environment, i.e. setpoint values computed by the EM are executed by the simulated PVBSS. The response of the environment is given by the unknown reward distribution $r_t \sim \mathcal{R}_a(\cdot|s_t, a_t)$ and state transition probability distribution $s_{t+1} \sim \mathcal{P}_a(\cdot \mid s_t, a_t)$. The general objective of a reinforcement learning algorithm is to derive an optimal policy $\pi^*(s_t)$, which maximizes the expected value of discounted rewards:

$$\pi^*(s_t) = \underset{\pi}{argmax}\, \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | \pi\right] \tag{1}$$

To address this policy optimization problem, the EM concept proposed in this paper uses the PPO algorithm, which shows fast convergence and learns better policies than other RL algorithms in a variety of continuous control problems [21], [22]. In recent years various RL algorithms have been applied to address the problem of optimal energy management in battery storage systems. Kuznetsova et al. [23] and Chenxiao Guan et al. [24] use Q-learning for the energy management of a battery storage system in a residential home with renewable power sources. Kim and Lim [25] apply the same algorithm for optimal power scheduling in a smart energy building. However, because the Q-learning algorithm only allows discrete states and actions, a discretization must be performed in each of these papers. As described in Ji et al. [26], the discretization of actions can degrade the performance of the EM and becomes unfeasible with higher dimensionality of the action space. Therefore, [26] use the continuous PPO algorithm for deriving a control policy for a micro grid management. However, in contrast to the investigations described in this paper, the temporal resolution of the simulation in [26] is much broader with one hour time intervals. Desportes et al. [27] apply another continuous RL algorithm, the Deep Deterministic Policy Gradient algorithm [28], for optimal power allocation in hybrid storage systems. In addition to the energy management of stationary battery storage systems, RL algorithms are also applied for other energy storage applications, such as thermal energy storage systems [29], residential homes with demand side management [30] or determination of optimal power flows in electric vehicles [31], [32].

C. NEURAL NETWORKS, PARTIAL OBSERVABILITY, AND OPTIMIZATION

In real-world EM applications, it is not feasible to capture the state variables s_t of the PVBSS completely. Instead, the policy has to derive relationships between incomplete observations $o_t \in \Omega, o_t \sim \mathcal{O}(\cdot|s_t)$, in this case the measured PV and load power as well as the SOC, and the underlying, hidden state s_t, for example weather conditions or load scenarios. The previously formulated MDP describing the EM of the PVBSS can thus be extended to a partially observable Markov decision problem (POMDP) $(\mathcal{S}, \mathcal{A}, \mathcal{P}_a, \mathcal{R}_a, \gamma, \mathcal{O}, \Omega)$. To estimate the hidden state s_t, a set of n past observations $o_{\leq t} = (o_{t-n} \ldots o_{t-1}, o_t)$ can be used as input for the policy $a_t \sim \pi(\cdot|o_{\leq t})$ [33].

The PPO algorithm assumes a differentiable function for the policy, parameterized by the vector θ. In the context of POMDPs, recurrent neural networks, such as LSTM networks, have been established for this purpose [34], [35]. LSTM networks retain an inner state, which allows the contextualization of observations $o_{\leq t}$ over several time steps and allows a better estimation of the hidden state s_t. It has been repeatedly shown that LSTM networks can represent highly complex policies for POMDPs. [36], [37], [38], [39], [40].

In training, the policy network is first initialized with a random parameter vector θ. Trajectories are then sampled by executing actions a_t under the current stochastic policy π_θ:

$$a_t \sim \pi_\theta\left(\cdot \mid o_{\leq t}, \theta\right) \tag{2}$$

The resulting state transitions $(o_{\leq t}, a_t, r_t, o_{\leq t+1})$ are stored for the subsequent optimization of the parameter vector θ by the PPO algorithm. A second neural network is used to estimate the state value function $V(o_{\leq t})$, which is defined as the expected sum of discounted rewards under the current policy:

$$V\left(o_{\leq t}\right) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | o_{\leq t}\right] \tag{3}$$

Based on the state value function $V(o_{\leq t})$ estimated by the second neural network, the advantage values $\hat{A}_t$ are calculated. They are defined as difference of the expected reward and the discounted reward received under the current policy:

$$\begin{aligned}\hat{A}_t = &-V\left(o_{\leq t}\right) + r_t + \gamma r_{t+1} \\ &+ \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V\left(o_{\leq T}\right)\end{aligned} \tag{4}$$

In the course of training, the prediction error of the state value function is minimized with regard to the objective function $L^V(\theta)$:

$$L^V(\theta) = \left(V_\theta\left(o_{\leq t}\right) - r_t + \gamma V_\theta\left(o_{\leq t+1}\right)\right)^2 \tag{5}$$

The objective function for the parameter optimization of the policy network is the PPO-specific surrogate function:

$$L^{\pi}(\theta) = \mathbb{E}_t\left[min\left(r(\theta), clip\left(r(\theta), 1-\epsilon, 1+\epsilon\right)\right)\hat{A}_t\right] \tag{6}$$

The derivation of this objective function is beyond the scope of this paper and can be found in [21]. Practically, actions resulting in a positive advantage are made more likely in the stochastic distribution of the policy. The PPO-specific objective function allows the collected state transitions to be used for multiple consecutive parameter updates without divergence of the optimization. This increases sample efficiency and reduces training time.

The combined cost function $L(\theta)$ for the policy and state value functions is obtained from weighted terms of respective cost functions $L^{V}(\theta)$ and $L^{\pi}(\theta)$. Furthermore, a weighted term for maximizing the entropy of the policy $H\left(\pi_\theta(\cdot|o_{\leq t}, \theta)\right)$ is introduced. This encourages explorative behavior of the RL-agent and helps to prevent convergence to a suboptimal policy:

$$L(\theta) = -L^{\pi}(\theta) + \beta_{Value}L^{V}(\theta) - \beta_{\mathcal{H}}\mathcal{H}\left(\pi_\theta\left(\cdot \mid o_{\leq t}, \theta\right)\right) \tag{7}$$

III. MODELING

The simulated PVBSS consists of a PV system with an inverter, accumulated loads and a battery storage with another inverter. As shown in figure 1, the system components are coupled via the AC grid of the residential building, which in turn is connected to the general power grid. The modeling parameters of the PVBSS can be found in appendix A.1.

The PVBSS is simulated in discrete time intervals of $\Delta t = 1\text{min}$. The length of the control interval is set to $\Delta t_{Control} = 15\text{min}$, similar to other studies [41]. To simplify the implementation, $\Delta t_{Control}$ is always a multiple of the simulation interval Δt. Therefore the modeling parameter $n_{SPA} = 15$ is introduced:

$$\Delta t_{Control} = n_{SPA}\Delta t \tag{8}$$

The PVBSS is implemented in an object-oriented form in the Python programming language. The components of the PVBSS, such as the grid connection, the electrical load or the PV system are abstracted by their own classes. In addition, a wrapper class called Environment is implemented, which contains functions for the execution of a simulation step and the calculation of the state and reward. It serves as an interface class between the simulated PVBSS and the RL algorithm.

A. PV SYSTEM, LOADS, AND GRID CONNECTION

In order to represent a wide range of realistic weather conditions and load scenarios, the simulation of PVBSSs is usually carried out based on historical data of the PV power, load power and electricity prices [23], [24], [25]. Here, the PV system is modeled by a measured time series of the global radiation data G_t, which was recorded by the Chair of Meteorology of TU Dresden at the Tharandt weather station (see Figure 2). This dataset covers the period from 2015 to 2019 with a temporal resolution of ten minutes. The output power of the PV generator is calculated by the surface area of the PV generator A_{PV} and the efficiency coefficient η_{PV}, which simplifies the module efficiency, shading and temperature-dependent effects of the PV system. Therefore, the modeling of the PV system only aims to represent the variance between different seasons and weather conditions:

$$P_{PV,t} = G_t A_{PV}\eta_{PV} \tag{9}$$

$$E_{PV,t} = P_{PV,t}\Delta t \tag{10}$$

The load power is modelled from a dataset from HTW Berlin, which contains measurements of 74 households over the course of the year 2010 at a temporal resolution of one minute (see Figure 2). The measurement series of the 74 households are clustered by their annual total energy and variance using the k-means algorithm. The cluster with a mean total energy of 4,200 kWh including 22 households is selected as the load power dataset.

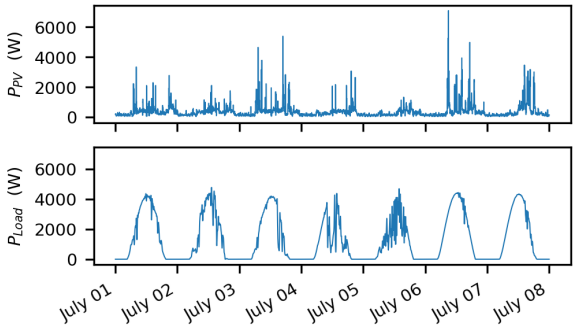

FIGURE 2. Time series of PV power P_{PV} and load power P_{Load} over a period of seven days from a generated training episode.

A constant electricity price of $p_{GD} = 0.32\ €\text{kWh}^{-1}$ and a constant feed-in remuneration of $p_{GF} = 0.08\ €\text{kWh}^{-1}$ are assumed for the grid connection. Energy that is exchanged with the distribution grid is measured by the meter E_{GD} for energy drawn from the grid and E_{GF} for energy fed into the grid. A feed-in limit of 50 % of the PV systems peak power $P_{GF,max} = 2.5\text{kW}$ is implemented. Power flow exceeding this limit in feed-in direction leads to the curtailment of the PV system and its integral is virtually metered as E_{CL}. This feed-in limit is based on the 50 % limit of the German KfW subsidy program for private PV systems [20], [42], which has been suspended as of 2023 but is still useful to consider because it reduces the load put onto the low voltage grid.

The time series data of the global radiation and load power is split into segments of one year length, yielding five years of global radiation data and 22 years of load power data, each with a temporal resolution of $\Delta t = 1\text{min}$. By permuting these time series, 110 episodes are generated. Of these, 12 are randomly excluded from the set of training episodes and retained as evaluation episodes.

B. ENERGY STORAGE

The EM provides the grid power setpoint for the PVBSS $P_{Grid,set,t}$. Throughout the control interval $\Delta t_{Control}$, the inter-

nal controller of the battery storage adjusts its battery power setpoint $P_{BAC,set,t}$ for charging and discharging according to the current residual power between the PV and load power $P_{Res,t}$:

$$P_{BAC,set,t} = P_{Grid,set,t} - P_{Res,t} \tag{11}$$

The efficiency of the battery inverter as well as the losses in the battery cells of the storage are represented by a constant, combined efficiency η_{Bat}, which is applied both when charging and discharging:

$$P_{BDC,set,t} = \eta_{Bat} P_{BAC,set,t} \tag{12}$$

$$P_{BDC,act,t} = \min\left(P_{BDC,set,t}, \frac{(1 - SOC_t)\, C_{Bat}}{\Delta t}\right) \tag{13}$$

$$P_{BAC,act,t} = \frac{P_{BDC,act,t}}{\eta_{Bat}} \tag{14}$$

Discharging the battery storage is similar to charging according to the following calculation:

$$P_{BDC,set,t} = \frac{P_{BAC,set,t}}{\eta_{Bat}} \tag{15}$$

$$P_{BDC,act,t} = \max\left(P_{BDC,set,t}, \frac{(-SOC_t)\, C_{Bat}}{\Delta t}\right) \tag{16}$$

$$P_{BAC,act,t} = P_{BDC,act,t} \eta_{Bat} \tag{17}$$

C. REINFORCEMENT LEARNING INTERFACE

The interaction of the RL agent with the simulated PVBSS is described by three fundamental values: the state s_t – or in terms of the POMDP formulation, the observation $o_{\leq t}$, the action a_t and the reward r_t. Kuznetsova et al. [23], Chenxiao Guan et al. [24] and Kim and Lim [25] includes the values of the load power P_L, the PV power P_{PV} and the battery state of charge SOC in the state vector s_t, which is the input of the RL agent. [24] include the current electricity price p_{GD} additionally. In all of these investigations, the action is the battery power setpoint $P_{BAC,set}$ and the reward is a function of the cost of energy drawn from the grid and energy fed into the grid $r_t = f(E_{GD,t}, E_{GF,t})$.

1) OBSERVATIONS

In the considered PVBSS, a single observation o_t consists of three values: the current state of charge of the storage in percent SOC_t, the cumulative energy consumption over the last time interval $E_{L,t}$ and the energy generated by the PV generator $E_{PV,t}$. A series of n_{Tail} previous observations is concatenated as an observation matrix $o_{\leq t}$ and returned to the RL algorithm.

$$o_t = \begin{bmatrix} SOC_t \\ E_{L,t} \\ E_{PV,t} \end{bmatrix} \tag{18}$$

$$o_{\leq t} = \left[o_{t-n_{Tail}} \cdots o_{t-1}, o_t\right] \tag{19}$$

2) ACTIONS

The action vector a_t contains the setpoints for the environment, i.e. the grid power setpoint of the PVBSS $P_{Grid,set,t}$. During training actions are sampled from the stochastic policy $\pi(o_{\leq t})$. The outputs of the policy are limited to the value range $a_t \in [-1, 1]$ for technical reasons. The actions a_t are therefore interpreted as the grid power setpoint normalized by the maximum charging or discharging power of the battery inverter $P_{BAC,max}$.

$$a_t = \left[\frac{P_{Grid,set,t}}{P_{BAC,max}}\right] \tag{20}$$

3) REWARD

The rewards r_t are calculated based on measurement values of the PVBSS and are then return to the RL algorithm. They are formulated as the revenue of the energy fed into the grid, minus the cost of the energy drawn from the grid. Therefore, the minimization of energy costs is defined as objective for the policy optimization. Implicitly, this requires the maximization of self-sufficiency and the minimization of PV curtailment.

$$r_t = E_{GF,t} p_{GF,t} - E_{GD,t} p_{GD,t} \tag{21}$$

IV. REINFORCEMENT LEARNING BASED ENERGY MANAGEMENT

A. NEURAL NETWORK TOPOLOGY

As mentioned in section A.3 the hidden system state variables s_t cannot be captured directly. Instead, a part of the neural network topology, the state encoder, is used to extract features from a matrix of past observations $o_{\leq t}$. The state encoder is shared between the policy network and the state value network and is represented by a LSTM network followed by a dense layer (see Figure 3).

Linked to the state encoder, the policy also has a separate part of the neural network for computing actions a_t. Similarly, the state value function has a separate part for computing the state values $V(o_{\leq t})$. Both of these separate neural networks consist of three dense layers. The output of the policy is squashed by the tanh activation function to the value range of $a_t \in [-1, 1]$. During inference, these output values are directly used as action a_t. During training however, the output values serve as the mean value for a parametrized truncated normal distribution, which is needed for the exploration of the state space S. Recent literature suggests that bounded probability distributions are better suited for sampling actions in RL than unbounded ones [43].

The neural networks of the policy and state value function as well as the common state encoder are implemented in the Python programming language using the Tensorflow library.

B. SIMULATION-BASED POLICY TRAINING

During training, actions a_t are sampled from the truncated normal distribution of the current policy. Actions a_t are passed to the simulation of the PVBSS and are scaled by the value of the maximum power of the battery inverter $P_{BAC,max}$. A number of n_{SPA} time steps are simulated, and the relevant energy quantities $E_{PV,t}$, $E_{L,t}$, $E_{GD,t}$, $E_{GF,t}$ and

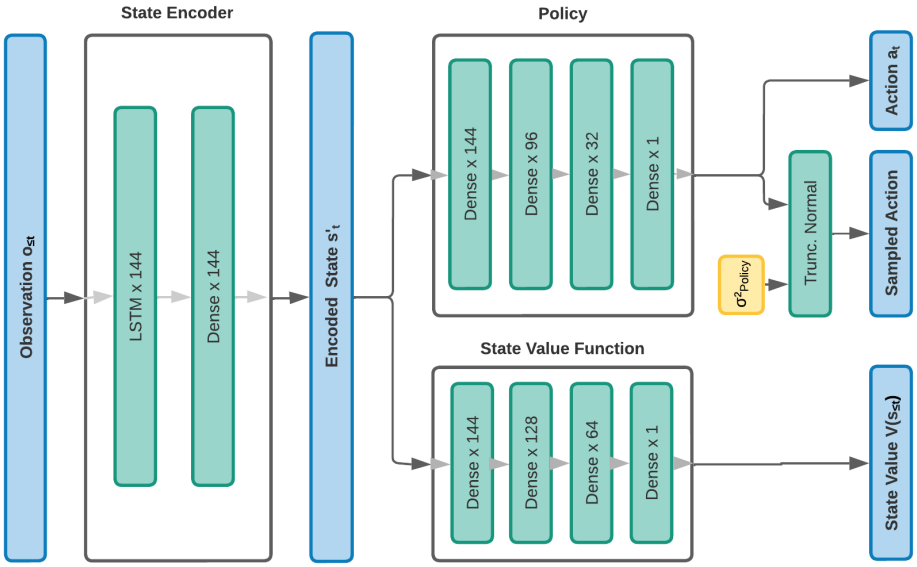

FIGURE 3. Neural network topology of the policy and of the state value function with a shared LSTM and a dense layer used as state encoder as well as separate dense layers for the state value function and the policy.

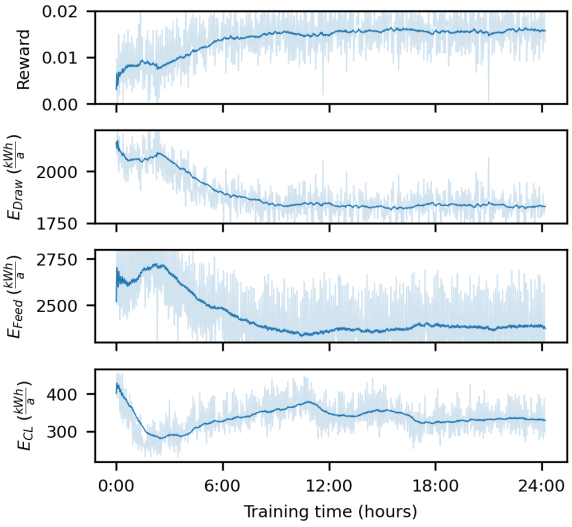

FIGURE 4. Convergence behavior of the received reward, the energy drawn from the grid E_{GD}, the energy fed into the grid E_{GF} and the PV curtailment losses E_{CL} over a year for a training period of 24 hours.

$E_{CL,t}$ are measured in the process. From the measured energy quantities $E_{PV,t}$ and $E_{L,t}$ as well as the state of charge of the battery storage SOC_t the next observation matrix $o_{\leq t+1}$ and the reward signal $r_t = f(E_{GD,t}, E_{GF,t})$ is calculated. The EM passes these two values back to the PPO algorithm, which stores these state transitions $(o_{\leq t}, a_t, r_t, o_{\leq t+1})$ for the subsequent optimization of the neural network parameters. After the optimization these state transitions are discarded and new ones are collected under the optimized policy (see Equation 6) [21].

Concurrent with the training, the current deterministic policy is evaluated. This allows tracking the state of progress without the noise of stochastic exploration. As shown in figure 4, over the course of the training, the following improvements can be observed in the simulated PVBSS: the average reward over the evaluation episodes $T^{-1} \cdot \sum_{t=0}^{T} r_t$ increases until convergence is achieved, indicating the policy has improved with regard to the optimization objective of minimizing the overall energy costs, as defined in the reward function. The energy drawn from the distribution grid E_{GD} and the energy fed into the grid E_{GF} decrease, suggesting the degree of self-sufficiency has improved. Further, the PV curtailment loss E_{CL} has decreased.

C. HYPERPARAMETER STUDY

Variations around the default values of selected hyperparameter of the generic PPO implementation [44] are investigated to quantify and interpret their influence on the convergence of the RL algorithm and the converged policy [45]. These include the learning rate α, the discount factor γ, the generalized advantage estimate (GAE) factor λ, the entropy regularization β_H and the learning rate halving. The variations are selected in linear or logarithmic increments, based on the valid range of values of the respective hyperparameter. In total, 216 policies are trained in a full training run for all permutations of the values $\alpha \in \{3.3 \cdot 10^{-5}, 10^{-4}, 3.3 \cdot 10^{-4}\}$, $\gamma \in \{0.99, 0.995, 0.9975, 0.99875\}$, $\lambda \in \{0.7, 0.8, 0.9\}$, $\beta_H \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ and learning rate halving enabled Yes, No. These extensive computations are carried out on the High Performance Computing cluster of TU Dresden.

The exponential moving average (EWMA) of the reward with a decay of $\alpha_{EWMA} = 0.005$ is considered at different stages of the training: at the beginning $n_{Iter} = 6,000$, after a few hours $n_{Iter} = 25,000$ and at the end of the training $n_{Iter} = 60,000$. Larger values of the learning rate α cause faster adaptation of the neural networks thus a faster progress at the beginning of the training ($n_{Iter} = 6,000$, $n_{Iter} = 25,000$). At the end of the training ($n_{Iter} = 60,000$) however,

a wide scatter of the data points can be observed. This is the case because the training can diverge at higher learning rates. Therefore, a learning rate of $\alpha \leq 10^{-4}$ is recommended as a reference point for the energy management application considered in this paper.

The entropy regularization $\beta_{\mathcal{H}}$ provides a surprising result. The highest value of $\beta_{\mathcal{H}} = 0.01$ provides the best results and also shows the narrowest scatter of data points. Eight out of ten of the best regulation policies had the value of $\beta_{\mathcal{H}} = 0.01$. These results indicate that even greater values for the entropy regularization $\beta_{\mathcal{H}} > 0.01$ should be investigated.

The examination of the discount factor γ shows that very high values of $\gamma \in \{0.9975, 0.99875\}$ have a negative impact on the learning behavior and the final control policy. These results are plausible as the discount factor implies an effective horizon of future rewards considered for the policy optimization. This is derived from the limit of the infinite series of discount values:

$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma} \tag{22}$$

For the value $\gamma = 0.9975$ this results in an effective horizon of 400 time steps, for $\gamma = 0.99875$ of 800 time steps. Such long planning horizons are not useful for the considered energy management application and lead to an increase in the variance of the discounted rewards. An effective horizon of 100 to 200 time steps, which corresponds to a discounting factor of $0.99 \leq \gamma \leq 0.995$, seems to be a reasonable choice for the considered EM application.

The variation of the GAE factor λ has only a minor effect on the training compared to the variation of other hyperparameters. The scatter along the y-axis is significantly larger than the trend component at each stage of the training. Therefore, the default value for the GAE factor of $\lambda = 0.9$ chosen in [44] seems to be well suited for this problem.

Enabling learning rate halving decreases the learning rate whenever the RL algorithm gets stuck in a local reward maximum for more than 6000 iterations. This causes the parameter updates of the optimization to become increasingly smaller and forces the convergence of the algorithm. Especially towards the end of the training $n_{Iter} = 60000$ a much narrower scatter can be observed, indicating a more reliable performance of the PPO algorithm. For each of the ten best training runs rate halving had been enabled.

V. ENERGY MANAGEMENT EVALUATION

The EWMA of the reward in each of the 216 training runs (see A.3) is evaluated after the end of the training, i.e. after 48 hours of training time has elapsed. The five best policies with regard to this value are used to simulate the episodes of the evaluation dataset (see Appendix A.3). Policy selection after full knowledge of the evaluation results should be avoided, because it is not a realistic assumption for the online operation. The priority-based EM (PRIO) and the model-predictive EM (MPC) are used as references for the evaluation of the EMs learned by the RL algorithm (RL-∗).

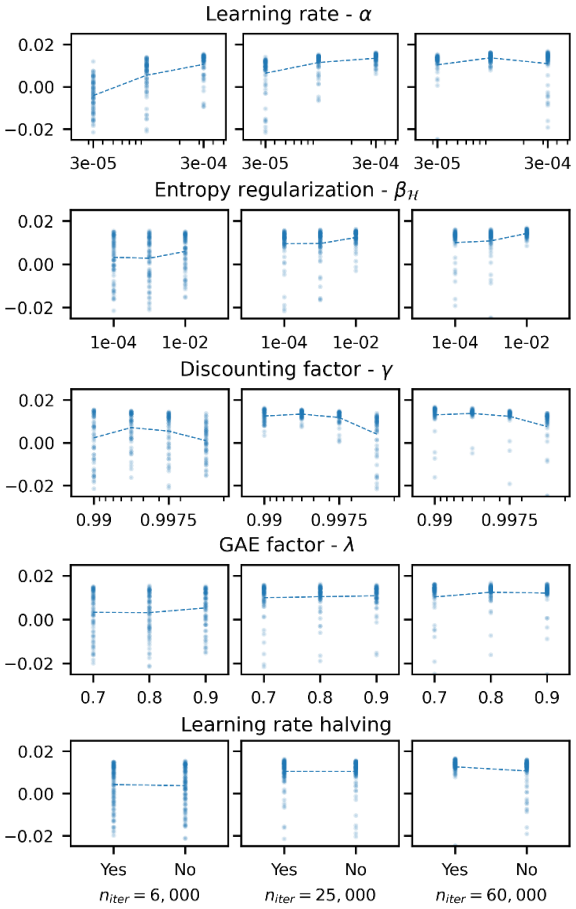

FIGURE 5. EWMA values of the rewards after 6,000, 25,000 and 60,000 training iterations for variation of the studied hyperparameters $\alpha, \beta_{\mathcal{H}}, \gamma, \lambda$ and learning rate halving.

As shown in figure 6, the RL-based EM, like the MPC-based EM, shifts the charging of the battery storage towards the end of the day and enables more energy being fed into the grid below the feed-in limit. This observation shows that the RL-based EM has learned a similar behavior to the MPC-based EM by using historical data of PV power and load power as input. In the MPC-based EM, this is achieved by formulating forecast models and by performing real-time optimization. Both add considerably to the complexity of the implementation as well as the computational power required, and are not needed for the RL-based EM. On the fourth day of the trajectory shown in Figure 6, it can be seen that the RL-based EM shifts the charging of the battery storage too far, resulting in the storage not being fully charged.

The EMs are evaluated based on the degree of self-sufficiency k_{SS}, the share of curtailment losses k_{CL}, the share of PV energy fed into the grid k_{FI} and the load-normalized specific energy costs k_{SEC}. These metrics are calculated from the annual trajectories of the evaluation episodes simulated under the respective EMs. The metrics are calculated from the accumulated energy quantities $*E$, which are the sum over the one-minute time intervals of the PVBSS simulation:

$$E_* = \sum_t E_{*,t} \tag{23}$$

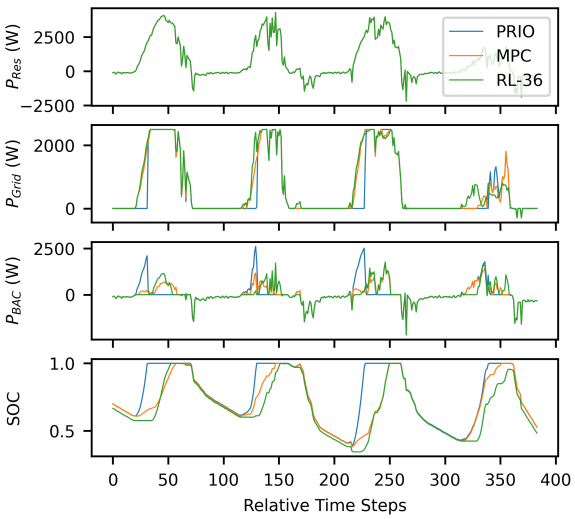

FIGURE 6. Trajectories of the residual power P_{Res}, power of the grid connection P_{Grid}, the AC power of the battery inverter $P_{BAC,act}$ and the state of charge for the PRIO, MPC and RL-36 energy managements over four summer days from evaluation episode 0.

$$k_{SS} = \frac{E_L - E_{GD}}{E_L} \tag{24}$$

$$k_{CL} = \frac{E_{CL}}{E_{PV}} \tag{25}$$

$$k_{FI} = \frac{E_{GF}}{E_{PV}} \tag{26}$$

$$k_{SEC} = \frac{p_{GD}E_{GD} - p_{GF}E_{GF}}{E_L} \tag{27}$$

The most relevant metric for the comparison of the EMs are the specific energy costs, since their minimization is set as the optimization objective defined by the reward function.

Table 1 shows the results of the simulation-based evaluation for an exemplary episode from the evaluation dataset for the five best RL-based EMs selected from the hyperparameter study (RL-∗) as well as the two EMs used as reference (PRIO, MPC). The lowest specific energy costs k_{SEC} in this evaluation episode are achieved by RL-36 with 7.09 CentkWh^{-1}. Several RL-based EMs can be learned that result in similar energy costs but that show different behaviors. For example, RL-13 feeds more energy into the grid than RL-103, but achieves a lower degree of self-sufficiency k_{SS}. Therefore, it can be concluded that despite learning an optimal strategy in terms of the reward function, the behavior can vary significantly. If a metric such as the degree of self-sufficiency k_{SS} shall be optimized explicitly, it must also be introduced in the reward function. Further, RL-36 achieves the lowest PV curtailment losses k_{CL} and the highest energy fed into the grid k_{FI}. The highest degree of self-sufficiency k_{SS} is achieved by PRIO as expected.

The statistics of all 12 evaluation episodes shown in figure 7 provide further evidence that the RL-based EMs achieve lower specific energy costs k_{SEC} than PRIO and similar energy costs to MPC. However, the degree of self-sufficiency k_{SS} and the share of energy fed into the grid k_{FI} highlight the different characteristics of the learned energy managements.

TABLE 1. Comparative metrics for the considered energy managements.

Energy management	k_{SS} (%)	k_{CL} (%)	k_{FI} (%)	k_{SEC} $\left(\frac{\text{Cent}}{\text{kWh}}\right)$
PRIO	**59.98**	9.87	45.39	7.54
MPC	59.51	5.70	49.95	7.16
RL-7	59.20	5.08	50.82	7.16
RL-13	59.18	5.09	50.82	7.17
RL-36	59.36	**4.78**	**50.99**	**7.09**
RL-43	59.43	5.18	50.54	7.12
RL-103	59.57	5.69	49.91	7.15

RL-103 shows the highest degree of self-sufficiency k_{SS} among the RL-based EMs in the distribution over all evaluation episodes but the lowest share of energy fed into the grid k_{FI}. RL-36 is the best EM with regard to the energy fed into the grid k_{FI} and the optimization objective of minimizing energy costs defined in the reward function.

VI. SUMMARY AND OUTLOOK

In this paper, an RL-based energy management concept using the Proximal Policy Optimization (PPO) algorithm was developed for the data-based learning of energy managements for a PV battery storage system (PVBSS). To address the partial observability of the PVBSS and the associated partially observable Markov decision process (POMDP) formulation of the energy management problem, Long Short-Term Memory (LSTM) networks were utilized to encode the systems state from a series of past observations. In an extensive hyperparameter study, 216 different energy managements were learned and the influences of selected hyperparameters on the convergence of the RL algorithm as well as the quality of the RL-based energy managements were analyzed.

Comparing the RL-based energy managements with the selected reference EMs over 12 evaluation episodes, similar or better energy costs were achieved compared to the state-of-the-art model-predictive EM. The trajectory of the state of charge of the RL-based EM is similar to that of the model-predictive EM, in that it shifts the charging of the battery storage towards the end of the day and thereby reduces curtailment losses and increases the amount of energy fed into the grid. This is achieved without the need for modeled forecasts of the PV or load power and without real-time optimization. Multiple EMs learned by the RL algorithm with different hyperparameter settings have achieved a similar optimum in terms of the energy costs. However, they showed different weightings of self-consumed PV energy and energy fed into the grid.

In future research, further optimization objectives shall be considered in the reward function. This includes efficiency- and aging-optimized operation of the battery storage as well as the consideration of the costs of the peak power drawn from the grid and time-variable energy prices. The PV system, the inverter as well as the battery storage shall be modeled by a more realistic loss behavior and semi-empirical battery

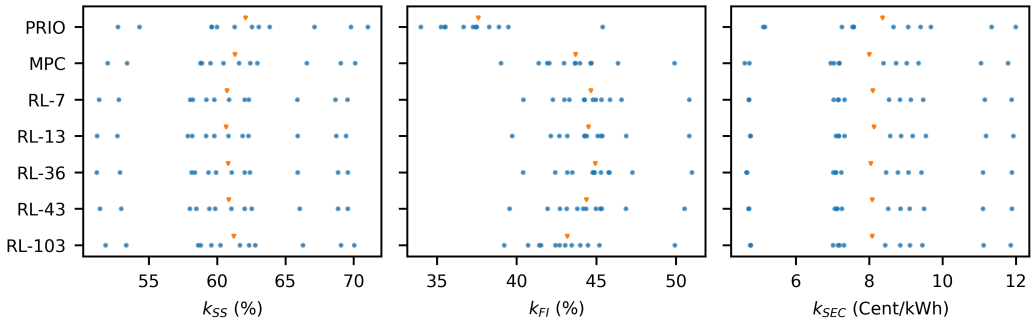

FIGURE 7. Statistics of the degree of self-sufficiency k_{SS}, the share of PV energy fed into the grid k_{FI} and the specific energy costs k_{SEC} over twelve evaluation episodes (blue) and their mean value (orange).

aging models shall be included in the simulations. For this purpose, efforts are currently being made to improve the implementation and to reduce the computation time in order to deal with the increased computational complexity.

To further improve the performance and robustness of the RL approach, various neural network topologies are investigated, e.g. the number of dense layers, the number of LSTM units and whether the state encoder is shared between the policy and the state value function. Additional options, such as the choice of the optimizer, the use of dropout and weight regularization will also be explored. The application of RL-based EMs for hybrid energy storage systems with more than one degree of freedom is currently conceptualized.

GLOSSARY

GENERAL ABBREVIATIONS

EWMA	Exponentially weighted moving average.
GAE	Generalized advantage estimate.
LSTM	Long Short-Term Memory.
MDP	Markov decision process.
MPC	Model-predictive control.
PV	Photovoltaic.
RL	Reinforcement learning.
POMDP	Partially observable Markov decision process.
PPO	Proximal Policy Optimization.
PRIO	Priority-based energy management.
PVBSS	PV battery storage system.
SOC	State of charge

Symbols

$\mathcal{A}$	Action space.
A_{PV}	PV module surface area.
$\hat{A}_t$	Advantage values.
a_t	Action.
$\cdot_{act}$	Actual value.
C_{Bat}	Battery capacity.
E_{GD}	Energy drawn from the grid.
E_{GF}	Energy fed into the grid.
E_L	Energy consumed by the loads.
E_{PV}	Energy converted by the PV generator.
E_{CL}	Energy of PV curtailment.
η_{Bat}	Battery efficiency.
η_{PV}	Effective PV efficiency.
$f(\cdot)$	Generic function.
G_t	Global radiation.
$\mathcal{H}(\cdot)$	Entropy of random distribution.
k_{SS}	Degree of self-sufficiency.
k_{SC}	Share of self-consumption.
k_{FI}	Share of energy fed into the grid.
k_{CL}	Curtailment losses.
k_{SEC}	Specific energy cost.
$L(\cdot)$	Loss function.
$\cdot_{max}$	Maximum value.
n_{SPA}	Number of simulation steps per action.
n_{Iter}	Number of training iterations.
$\mathcal{O}$	Observation space.
o_t	Observation.
$o_{\leq t}$	Matrix of past observations.
$\mathcal{P}_a$	State transition distribution.
Ω	Observation distribution.
p_{GD}	Electricity price.
p_{GF}	Feed-in remuneration.
P_{BAC}	Battery storage AC power.
P_{BDC}	Battery storage DC power.
P_G	Grid power.
P_L	Load power.
P_{PV}	PV generator power.
P_{Res}	Residual power.
$\pi(\cdot)$	Policy function.
$\mathcal{R}_a$	Reward distribution.
r_t	Reward.
$r(\cdot)$	Probability ratio.
$\mathcal{S}$	State space.
s_t	State.
$\cdot_{set}$	Setpoint.
Δt	Simulation time interval.
$\Delta t_{Control}$	Control tine interval.
T	Terminal time step.
θ	Parameter vector of the neural network.
$\cdot_t$	Value at time step t.
$V(\cdot)$	State value function.

HYPERPARAMETERS

see Appendix A.2

APPENDIX

A. MODELING PARAMETERS

TABLE 2.

Symbol	Value	Description
Δt	1 min	Simulation time interval
$\Delta t_{Control}$	15 min	Control time interval
n_{SPA}	15	Simulation steps per action
A_{PV}	50 m^2	PV surface area
η_{PV}	0.1	Effective PV efficiency
$P_{GF,max}$	2,500 W	Maximum feed-in power
$P_{BAC,max}$	4,000 W	Maximum battery inverter power
η_{Bat}	0.92	Battery storage efficiency
C_{Bat}	7 kWh	Battery storage capacity
p_{GD}	0.32 € kW^{-1} h^{-1}	Electricity price
p_{GF}	0.08 € kW^{-1} h^{-1}	Feed-in remuneration

B. HYPERPARAMETERS

TABLE 3.

Hyperparameter		Value
Batch size		4,096
Learning rate	α	$\{3.3 \cdot 10^{-5}, 10^{-4}, 3.3 \cdot 10^{-4}\}$
Discount factor	γ	$\{0.99, 0.995, 0.9975, 0.99875\}$
Weight value loss	β_{Value}	1.0
Weight entropy loss	$\beta_{\mathcal{H}}$	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
Tail length		96
Environment steps per iteration		256
Clipping factor	ϵ	0.2
GAE factor	λ	$\{0.7, 0.8, 0.9\}$
Initial log variance	$\log(\sigma^2_{init})$	-0.5
Training epochs	k_{Epochs}	6

C. EVALUATION EPISODES

TABLE 4.

i	Total PV energy (kWh)	Total load energy (kWh)
0	6,104	4,210
1	4,977	4,777
2	5,452	4,251
3	6,182	4,267
4	6,104	4,536
5	4,977	4,536
6	6,182	4,351
7	5,932	4,536
8	4,977	4,351
9	4,977	4,203
10	5,932	4,926
11	5,452	4,926

ACKNOWLEDGMENT

The authors would like to thank the Centre for Information Services and High Performance Computing (Zentrum für Informationsdienste und Hochleistungsrechnen—ZIH), TU Dresden, for providing its facilities for high throughput calculations and the Chair of Meteorology with TU Dresden for providing irradiation data for their simulations.

REFERENCES

[1] *Bundes-Klimaschutzgesetz: KSG*, Federal Ministry Environ., Nature Conservation Nucl. Saf., Germany, 2021.
[2] M. Sterner and I. Stadler, *Energiespeicher—Bedarf, Technologien, Integration*, 2nd ed. Berlin, Germany: Springer, 2017.
[3] P. Sterchele. (Feb. 2020). *Wege Zu Einem Klimaneutralen Energiesystem: Die Deutsche Energiewende Im Kontext Gesellschaftlicher Verhaltensweisen*. Freiburg im Breisgau. [Online]. Available: https://www.ise.fraunhofer.de/de/veroeffentlichungen/studien/wege-zu-einem-klimaneutralen-energiesystem.html
[4] J. Figgener, C. Hecht, D. Haberschusz, J. Bors, K. Gerd Spreuer, K.-P. Kairies, P. Stenzel, and D. U. Sauer, "The development of battery storage systems in Germany: A market review (status 2023)," 2022, *arXiv:2203.06762*.
[5] O. Schmidt, A. Hawkes, A. Gambhir, and I. Staffell, "The future cost of electrical energy storage based on experience rates," *Nature Energy*, vol. 2, no. 8, pp. 1–8, Jul. 2017, doi: 10.1038/nenergy.2017.110.
[6] X. N. Penisa, M. T. Castro, J. D. A. Pascasio, E. A. Esparcia, O. Schmidt, and J. D. Ocon, "Projecting the price of lithium-ion NMC battery packs using a multifactor learning curve model," *Energies*, vol. 13, no. 20, p. 5276, Oct. 2020, doi: 10.3390/en13205276.
[7] L. Mauler, F. Duffner, W. G. Zeier, and J. Leker, "Battery cost forecasting: A review of methods and results with an outlook to 2050," *Energy Environ. Sci.*, vol. 14, no. 9, pp. 4712–4739, Sep. 2021, doi: 10.1039/D1EE01530C.
[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. London, U.K.: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/
[10] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2019, doi: 10.17775/CSEEJPES.2019.00920.
[11] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.
[12] M. Bottiger, "Multikriteriell optimierendes betriebsfuhrungsverfahren fur PV-batteriespeichersysteme," Ph.D. dissertation, Dept. Mech. Sci. Eng., Inst. Power Eng., Chair Energy Storage Syst., TU Dresden, Dresden, Germany, 2019. [Online]. Available: https://katalog.slub-dresden.de/id/0-1724818279/#detail
[13] T. Weitzel and C. H. Glock, "Energy management for stationary electric energy storage systems: A systematic literature review," *Eur. J. Oper. Res.*, vol. 264, no. 2, pp. 582–606, Jan. 2018, doi: 10.1016/j.ejor.2017.06.052.
[14] M. Sterner, M. Thema, F. Eckert, and F. Bauer. *Der Positive Beitrag Dezentraler Batteriespeicher Für Eine Stabile Stromversorgung Kurzstudie Im Auftrag Des BEE E.V. Und Der Hannover Messe*. Accessed: Sep. 16, 2021. [Online]. Available: https://www.researchgate.net/publication/312951377_Der_positive_Beitrag_dezentraler_Batteriespeicher_fur_eine_stabile_Stromversorgung_Kurzstudie_im_Auftrag_des_BEE_eV_und_der_Hannover_Messe
[15] R. Gelleschus, M. Böttiger, and T. Bocklisch, "Optimization-based control concept with feed-in and demand peak shaving for a PV battery heat pump heat storage system," *Energies*, vol. 12, no. 11, p. 2098, Jun. 2019, doi: 10.3390/en12112098.
[16] T. Bocklisch, "Hybrid energy storage approach for renewable energy applications," *J. Energy Storage*, vol. 8, pp. 311–319, Nov. 2016, doi: 10.1016/j.est.2016.01.004.
[17] A. Nottrott, J. Kleissl, and B. Washom, "Energy dispatch schedule optimization and cost benefit analysis for grid-connected, photovoltaic-battery storage systems," *Renew Energy*, vol. 55, pp. 230–240, Jul. 2013, doi: 10.1016/j.renene.2012.12.036.
[18] M. Kersic, T. Bocklisch, M. Böttiger, and L. Gerlach, "Coordination mechanism for PV battery systems with local optimizing energy management," *Energies*, vol. 13, no. 3, p. 611, Jan. 2020, doi: 10.3390/en13030611.
[19] J. Moshövel, K.-P. Kairies, D. Magnor, M. Leuthold, M. Bost, S. Gährs, E. Szczechowicz, M. Cramer, and D. U. Sauer, "Analysis of the maximal possible grid relief from PV-peak-power impacts by using storage systems for increased self-consumption," *Appl. Energy*, vol. 137, pp. 567–575, Jan. 2015, doi: 10.1016/j.apenergy.2014.07.021.

[20] J. Bergner, J. Weniger, T. Tjaden, and V. Quaschning, "Feed-in power limitation of grid-connected PV battery systems with autonomous forecast-based operation strategies," in *Proc. 29th Eur. Photovoltaic Sol. Energy Conf. Exhib.*, 2014, pp. 2363–2370, doi: 10.4229/EUPVSEC20142014-5CO.15.1.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[22] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," 2017, *arXiv:1710.02298*.

[23] E. Kuznetsova, Y.-F. Li, C. Ruiz, E. Zio, G. Ault, and K. Bell, "Reinforcement learning for microgrid energy management," *Energy*, vol. 59, pp. 133–146, Sep. 2013, doi: 10.1016/j.energy.2013.05.060.

[24] C. Guan, Y. Wang, X. Lin, S. Nazarian, and M. Pedram, "Reinforcement learning-based control of residential energy storage systems for electric bill minimization," in *Proc. 12th Annu. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2015, pp. 637–642. [Online]. Available: http://ieeexplore.ieee.org/document/7158054/

[25] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, no. 8, p. 2010, Aug. 2018, doi: 10.3390/en11082010.

[26] Y. Ji, J. Wang, J. Xu, and D. Li, "Data-driven online energy scheduling of a microgrid based on deep reinforcement learning," *Energies*, vol. 14, no. 8, p. 2120, Apr. 2021, doi: 10.3390/en14082120.

[27] L. Desportes, I. Fijalkow, and P. Andry, "Deep reinforcement learning for hybrid energy storage systems: Balancing lead and hydrogen storage," *Energies*, vol. 14, no. 15, p. 4706, Aug. 2021, doi: 10.3390/en14154706.

[28] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[29] G. Henze and J. Schoenmann, "Evaluation of reinforcement learning control for thermal energy storage systems," *HVAC R Res.*, vol. 9, no. 3, pp. 259–275, Jul. 2003, doi: 10.1080/10789669.2003.10391069.

[30] S. Lee and D.-H. Choi, "Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, energy storage system, and home appliances," *Sensors*, vol. 19, no. 18, p. 3937, Sep. 2019, doi: 10.3390/s19183937.

[31] M. Wieczorek and M. Lewandowski, "A mathematical representation of an energy management strategy for hybrid energy storage system in electric vehicle and real time optimization using a genetic algorithm," *Appl. Energy*, vol. 192, pp. 222–233, Apr. 2017, doi: 10.1016/j.apenergy.2017.02.022.

[32] Y. L. Murphey, J. Park, Z. Chen, M. L. Kuang, M. A. Masrur, and A. M. Phillips, "Intelligent hybrid vehicle power control—Part I: Machine learning of optimal vehicle power," *IEEE Trans. Veh. Technol.*, vol. 61, no. 8, pp. 3519–3530, Oct. 2012, doi: 10.1109/TVT.2012.2206064.

[33] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson, "Deep variational reinforcement learning for POMDPs," 2018, *arXiv:1806.02426*.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[35] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," 2015, *arXiv:1507.06527*.

[36] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[37] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, and J. Oh, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[38] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," 2017, *arXiv:1712.01815*.

[39] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," 2016, *arXiv:1610.00633*.

[40] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," 2017, *arXiv:1707.01495*.

[41] J. Bergner, J. Weniger, and T. Tjaden. (2016). *PVprog-Algorithmus*. Dokumentation, HTW Berlin v1.1. Accessed: Jul. 5, 2020. [Online]. Available: https://fs-cloud.f1.htw-berlin.de/s/0DauTvabH2yCNe0?path=%2FVersion%201.1#pdfviewer

[42] J. Weniger, J. Bergner, T. Tjaden, and V. Quaschning, "50%-Studie: Effekte der 50%-einspeisebegrenzung des KfW-forderprogramms für photovoltaik-speichersysteme," Forschungsgruppe Solarspeichersysteme, HTW Berlin, 2016, doi: 10.13140/RG.2.2.34064.81929.

[43] P.-W. Chou, D. Maturana, and S. Scherer, "Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 834–843. [Online]. Available: http://proceedings.mlr.press/v70/chou17a.html

[44] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: https://jmlr.org/papers/v22/20-1364.html

[45] M. Andrychowicz, A. Raichuk, P. Stanczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, S. Gelly, and O. Bachem, "What matters in on-policy reinforcement learning? A large-scale empirical study," 2020, *arXiv:2006.05990*.

FLORUS HÄRTEL received the Dipl.-Ing. degree in mechanical engineering from Technische Universität Dresden, in 2020. He is currently pursuing the Ph.D. degree at this chair, continuing, and deepening the research efforts of his former Diploma thesis. He wrote his Diploma thesis with the Chair of Energy Storage Systems, Technische Universität Dresden, with a topic focusing on reinforcement learning and PV battery storage systems. His research interests include applied reinforcement learning for energy management strategies for battery storage systems and applied machine learning for different problems in the field of energy storage, such as time series predictions and uncertainty modeling.

THILO BOCKLISCH received the Diploma and Ph.D. degrees in electrical engineering from the Chemnitz University of Technology, Germany, in 2003 and 2009, respectively. From 2003 to 2008, he was a Research Assistant at the Chair of Electrical Machines and Drives, Chemnitz University of Technology, Germany. From 2008 to 2013, he was the Head of the Interdisciplinary Research Group "Intelligent Decentralized Energy Storage Systems," and from 2009 to 2016, he was the Head of the Research Group "Multi-Storage Hybrid Systems," Chair of Power Systems and High-Voltage Engineering, Chemnitz University of Technology. In 2016, he was appointed as an University Professor and the Head of the Chair of Energy Storage Systems, Technische Universität Dresden, Germany. He is the author of more than 50 scientific articles. His research interests include design, modeling, control and applications of hybrid energy storage systems, modeling, simulation and optimization of energy storage and energy conversion devices, sustainable energy supply concepts for power, and heat and transport, including sector coupling, prediction, and classification of energy time series.

He is the Chairperson of the Annual Energy Storage Symposium with Technische Universität Dresden. He received the Award of Dresdner Gesprächskreis der Wirtschaft und der Wissenschaft e.V. for outstanding results in his Ph.D. thesis, in 2011, and the Best Poster Award from the 8th International Renewable Energy Storage Conference, in 2013.