Reinforcement Learning for Residential Heat Pump Operation

Simon Schmitz[a,*], Karoline Brucke[b], Pranay Kasturi[c], Esmail Ansari[d], Peter Klement[b]

[a]*DLR-Institute for Software Technology, Linder Höhe , Cologne, 51147, Germany*
[b]*DLR-Institute of Networked Energy Systems, Carl-von-Ossietzky-Straße 15, Oldenburg, 26129, Germany*
[c]*Carl von Ossietzky University Oldenburg, Ammerländer Heerstraße 114-118, Oldenburg, 26129, Germany*
[d]*Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM, Wiener Straße 12, Bremen, 28359, Germany*

Abstract

Residential space heating accounted for approximately 19% of the overall energy consumption of Germany in 2021. Therefore, the efficient operation of electrified heating systems is of major importance for the energy transition. We apply a reinforcement learning approach for operating a district heat pump and compare results with a classic rule-based approach. No building model is required in our study but only basic parameters of the hot water tank along with demand and ambient temperature data which all is easily attainable. Additionally, the environment is designed in a way that the residents living comfort is never compromised which maximizes applicability in real world buildings. The agent is able to exploit variable electricity prices and the flexibility of the hot water tank in such a way, that up to 35% of energy costs could be saved. Additionally, depending on the agent's settings, only 23% to 41% of the heat pump's nominal power installed according to current standards was used. The robustness of the approach is shown by running ten independent training and testing cycles for all setups with reproducible results. The importance of demand forecasts is evaluated by testing different observation spaces of the RL agent. Even if the agent has no demand information at all, costs savings still are 25%.

*Corresponding author
Email address: Simon.Schmitz@dlr.de (Simon Schmitz)

Keywords: Reinforcement learning, heat pump operation, district heating, demand forecast, operation under uncertainty

1. Introduction

The residential sector caused approximately 28% of the total energy consumption in Germany in 2021 [1] while space heating accounted for more than two-thirds of this [2]. Therefore, the ongoing electrification of the residential heat sector with e.g. heat pumps (HP) provides challenges (e.g. increased over all electricity consumption) but also opportunities (e.g. demand response) for the overall energy system [3]. Providing flexibility on a decentral level will be important to ensure energy system stability especially in distribution grids. Especially, electrified heating systems offer a lot of flexibility due to the inherent inertia of most heating systems and sector coupling opportunities. According to [4], especially hot water tanks as heat storage are one of the most influential residential flexibility bearing devices and can be easily combined with HPs to electrify heat demands. But, two main things are important in harnessing this flexibility potential: High quality demand and generation predictions as well as intelligent control and operational management mechanisms based on the provided predictions [5].

While there is extensive research on carrying out demand and generation predictions, the operation of real world HPs is still rather simple [6, 7]. Currently, they are mostly operated based on classical control engineering using PID controllers or naive rule-based approaches which are not able to follow complex variable objectives or steering signals [7].

A technique called Reinforcement Learning (RL) is a promising approach for more sophisticated operational management of HPs being able to take into account the complexity of the respective heat system without requiring extensive model building as for Model Predictive Control (MPC) [8]. RL is a branch of machine learning that focuses on teaching agents to make optimal decisions in dynamic environments [9]. RL agents are able to take actions and learn from the resulting response of the environment, receiving rewards or penalties. RL has shown promise in solving complex problems in the energy context [10].

Applying RL to control and operational problems in residential heating is a rising research topic. RL has been applied to a variety of heat control problems mostly in the residential sector [4, 11] but also in commercial or

office buildings [12, 13]. While many publications consider only space heating, some also look at domestic hot water provided by HPs combined with hot water tanks like [14, 15]. Typically, RL can save 10% of energy costs when applied to the operation of HVAC systems and 20% for water heaters compared to rule-based approaches [10]. This has been demonstrated in various simulation studies [16, 11, 17]. Some publications already deploy their RL approaches into real world systems. In [18], the authors train an agent off-site on measured data and after an on-site training phase the RL agent takes over the HP control for domestic hot water in a real building. But generally, according to [8] RL for building control is still in research state with only limited applications in real world buildings (11% of studies).

A crucial step in designing a RL approach for building control is the selection of a meaningful reward function. This is mandatory to give the agent the needed feedback to optimize its behavior. For space heating control, this reward function is often designed using the internal room temperature together with comfort bands of the residents. This approach requires a building model which is able to interact with the inputs of the agent [19]. Most studies consider a temperature or comfort band of the residents somewhere between 19°C and 24°C for space heating like [20] or 24°C to 28°C for cooling like [12]. But during training and also sometimes during testing, the RL agent compromises living comfort. Additionally, this approach needs extensive expert knowledge creating the building models. Both factors decrease acceptance and deployment of the approach into real world systems [8].

A second important factor is the observation space of the RL agent, that is, the information that the agent gets about the current or future environment. In [13], the authors give the RL agent electricity price and weather forecasts and in [21], the authors investigate the impact of weather forecast quality on HP control. But they do not use RL but MPC. In general, most of the other works applying RL assume perfect foresight conditions or no foresight on the heat demand at all. But according to [5], it is important to combine and integrate forecasting into control problems for increased applicability in real world systems. RL is able to operate under forecast conditions as [22] shows for the RL-based operation of a hydrogen storage based on renewable generation forecasts. But despite the importance of the integration of forecasts into control, to the best of our knowledge, there is no publication applying RL to HP control for space heating based on heat demand predictions. Therefore, it is also unclear how important good external predictions of the heat demand are or whether RL is already able to find a sufficient

management strategy without such predictions.

The aim of this paper is the application of a RL approach to the operation of a residential district HP for space heating. Note, that heat demand from domestic hot water is not taken into account in this study. Our work includes the creation of a suitable but simple environment modeling of the heat network including a hot water tank as heat storage but without requiring a building model. The HP gets modeled using a temperature dependent coefficient-of-performance (COP) curve. Furthermore, the agent learns to operate under perfect foresight conditions as well as relying on demand predictions. The demand predictions are created using a recurrent neural network technique called Long-Short-Term-Memory (LSTM). Five years of simulated space heating demands with a granularity of 15 min are available. A rule-based approach is taken as benchmark for the results of the RL agent.

In this work, we present four main research contributions:

- Firstly, we demonstrate the operation of a HP using a RL approach working under perfect foresight conditions as well as forecast conditions for the respective heat demand. Doing so, we can quantify and evaluate the impact of demand uncertainty on the operation and respective costs of HPs using RL which has not been published before to the best of our knowledge.

- Secondly, in this work, no building model is required, since the building's inherent thermal inertia is assumed to be already decoded in the demand data of its residents. Instead, the only modeled heat storage is the installed hot water tank which is simulated with very basic parameters.

- The third contribution of this paper is that we carry out a RL-based operational management approach without ever compromising living comfort due to the environment's inherent condition that the heat demand of residents is met at all times. Flexibility will only be harnessed by exploiting the storage capacity of the modeled hot water tank. No building envelope and therefore also no indoor temperature is modeled. That approaching will also likely increase acceptance and adoption in real world heating systems significantly in the future.

- Lastly, we show the robustness and reproducibility of results by running ten independently trained agents on all our different tests and exam-

ining the means and standard deviations of all respective evaluation metrics.

This paper is structured as follows: First, we describe all data sources used for this paper in Section 2 which consist of heat demand data and weather data (Section 2.1) as well as historic variable electricity prices (Section 2.2). We follow, by extensively explaining the methodology and taken approaches in Section 3. That comprises a short introduction of the RL algorithm which was used in this work, followed by the description of the hot water tank model, the environment design, reward function design, demand forecast creation, describing the benchmark rule-based approach and finally presenting the evaluation metrics. Afterwards, we present the results in Section 4 by firstly examining effects of different RL agents on the district's energy costs and secondly by investigating the different learned operational strategies in more depth. This section is followed by a discussion of the results in Section 5 and is concluded by a summary and outlook in Section 6.

2. Data Sources

2.1. Simulated heat demand data and weather data

Note, that in this study we used simulated heat demand data but measured space heating demands are equally suited for applying our approaches. Heat demand due to domestic hot water is not taken into account. The historical heat demand profiles of a residential district were simulated using the software QuaSi [23]. The buildings under study were calibrated for a standard weather profile applying simplified cubatures and determining of the building material properties, in order to comply with the annual heat demand estimations according to the energy performance certificates of the buildings following DIN 4108 [24]. QuaSi simulates the buildings energetic behavior and thereby can create hourly or 15-minutes load profiles for space heating using a generic thermal building model based on EnergyPlus[25]. These calibrated models in QuaSi were then applied to generate the historical heat demand profiles using the historical hourly weather data published by DWD [26] from 2017 to 2021 for the location of Bremen, Germany. The few sporadic missing weather data were closed utilizing interpolation techniques based on reasonable assumptions. As QuaSi can only process the weather data in TRY-format, the historical weather data were hence mapped to this format. The simulation approach was used in order to have an extensive

data set to work with and test on. In this study the simulated heat demand was used in a quarter-hourly resolution. An exemplary representation of the heat demand and the ambient temperature for the period from 01.01.2020 to 01.01.2021 can be seen in Figure 1. As can be seen the heat demand reaches values of up to 50 kWh in the winter months and vanishes completely in the summer period. The ambient temperature reaches values of up to around 30 °C in summer and around −15 °C in winter.

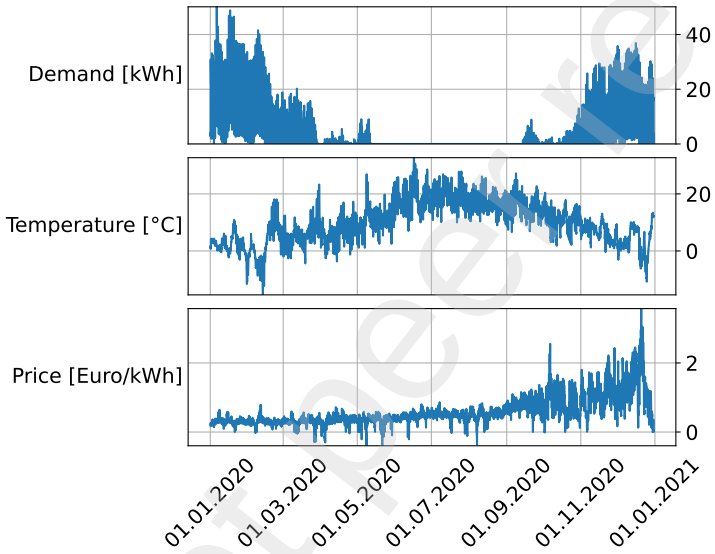

Figure 1: Heat demand, ambient temperature and electricity prices for a one year period.

2.2. Variable Electricity Prices

The variable electricity prices have been obtained from SMARD [27] which is the official provider of the electricity market data for Germany from the Federal Network Agency [28]. The original data source is ENTSO-E (European Network of Transmission System Operators for Electricity) [29]. Prices for the bidding zone *Germany/Luxembourg* were only available from October 1st 2018. Therefore, the prices from the bidding zone *Belgium* have been used for the period from January 1st 2017 to September 30th 2018. All prices are available in quarter-hourly resolution. Since the prices from SMARD are wholesale prices a scaling such that the mean of the prices is a typical consumption price of 0.3 Euro/kWh has been performed. Figure 1 shows the electricity prices for the period from 01.01.2020 to 01.01.2021. One

can see that the prices are rising in this period. Generally there is a lot of fluctuation with the lowest prices at just under 0 Euro/kWh and the highest prices with a little over 2 Euro/kWh.

3. Methodology

3.1. Reinforcement Learning Algorithm

The Proximal Policy Optimization (PPO) algorithm was used to train the RL agent in this work. The underlying theory is described in detail in [30]. PPO is an on-policy approach and does not require a model. It applies a "proximal" approach by introducing a clipping mechanism into the objective function. Thanks to this mechanism, the update of the strategy is within a certain range, preventing drastic changes that could lead to instability or divergence. One of PPO's notable strengths is its effectiveness in handling continuous action spaces. Traditional RL algorithms, such as Q-learning as described in [31], struggle with the high-dimensional and continuous nature of action spaces. PPO's policy-based approach, proximal updates and effectiveness in dealing with continuous action spaces make it particularly suitable for our task. For this work, the python implementation of PPO from Stable-Baselines3 has been used [32].

3.2. Heat storage model

The hot water tank model used for this work was developed for another project [33] and is openly available as a part of the mosaik-heatpump repository [34]. It is a multinode stratified thermal tank model, where the tank volume is divided into a specified number of layers (nodes) of equal volume, each characterized by a specific temperature. A traditional density distribution approach is adopted where the water flowing into the tank enters the layer that best matches its density (i.e., temperature). The model assumes that the fluid streams are fully mixed before leaving each of the layers and the flows between the layers follow the law of mass conservation. Heat transfer to the surrounding environment from the walls of the tank, and the heat transfer between the layers are considered.

The initial temperature profile inside the tank must be specified at the time of initialization of the model. For flows coming into the tank, both the temperature and flow rate should be specified. For the flows going out of the tank, only the flow rate should be specified, as the temperature is obtained

from the corresponding layer of the tank. The model ensures that the overall flow into and out of the tank is equal. The model then updates the temperatures of each layer based on the water flows through the specified connections, the heat transfer between the layers, and the heat transfer to the surrounding environment. The model has the functionality to flip the layers to ensure a negative temperature gradient from the top to the bottom of the tank. Finally, the model updates the connections with respect to the updated layer temperatures. For the flows going out of the tank, the temperature is updated. For the flows coming into the tank, the corresponding layer is updated.

3.3. Heat pump model

The HP is simulated via a linear regression that takes the inputs ambient temperature T_{amb} and water temperature T_{w} and predicts the COP. The linear regression is based on 18 measurements distributed in the range of -15°C and 20°C for the ambient temperature as well as 35°C and 55°C for the water temperature (see Table 1). The simulated COP value is used to calculate the thermal power P_{th} based on the chosen electric operational power P_{el} of the HP as follows.

$$P_{\mathrm{th}}(T_{\mathrm{amb}}, T_{\mathrm{w}}) = COP(T_{\mathrm{amb}}, T_{\mathrm{w}}) \cdot P_{\mathrm{el}} \tag{1}$$

P_{th} is used to determine the water flow $F_{\mathrm{HP,S}}$ from the HP to the heat storage (see Figure 2).

COP values	T_{w}						
T_{amb}	20	15	10	7	-2	-7	-15
35	5.61	–	4.45	4.21	3.75	3.07	2.56
45	–	–	–	3.44	3.11	2.59	2.21
50	4.58	3.66	–	3.11	2.82	2.37	–
55	3.59	–	–	2.80	–	2.29	–

Table 1: Measurements of the COP values for specific ambient temperatures T_{amb} and water temperatures T_{w}.

3.4. Environment Design

The environment described in this section has been built with Stable-Baselines3 [32] in combination with OpenAI's Gym library [35]. It consists

of a HP simulation as described in Section 3.3 as well as a heat storage simulation as described in Section 3.2. A schematic overview of the environment can be seen in Figure 3. The HP is of type air-to-water and has a nominal electrical power of 100 kW. The electrical power of the HP is continuously adjustable in a range from 0 to 100 kW and is chosen by the RL agent.

The installed nominal HP power results from the following: The district under consideration for simulation comprises approximately 100 residential living units and 7000 square-meters of living space at a space heating demand of 25-28 kWh/qm per annum. Taking into account best practices and security concerns for sizing heat pumps in the climate environment of northern Germany this would result in a nominal power of 200 kW including domestic hot water. Since domestic hot water is not considered in this work due to data availability and accounts for approximately half of the total heat demand only 100 kW of nominal power are assumed here.

The heat storage for the whole district is one hot water tank with a height of $H = 5\,\mathrm{m}$ and a diameter of $D = 4\,\mathrm{m}$. These dimensions result in a volume of almost 63 000 l. This single hot water tank is an aggregation of the multiple smaller hot water tanks that would actually be installed in the considered district. The hot water tank has connectors at $h_{\mathrm{HP,S}} = h_{\mathrm{S,D}} = 4.999\,\mathrm{m}$ for the hot water and connectors at $h_{\mathrm{S,HP}} = h_{\mathrm{D,S}} = 0.001\,\mathrm{m}$ for the cool water.

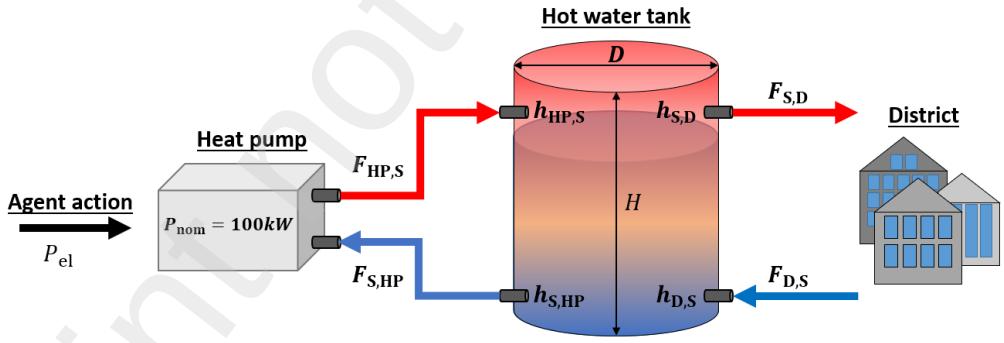

Figure 2: Schematic representation of the used district heating model including the a hot water tank for heat storage

The water flow $F_{\mathrm{HP,S}}$ between the HP and the hot water tank is calculated by

$$F_{\mathrm{HP,S}} = \frac{P_{\mathrm{th}}}{c_{\mathrm{water}} \cdot \Delta_{\mathrm{T,HP}}} \tag{2}$$

where c_{water} denotes the specific heat capacity of water and $\Delta_{\text{T,HP}}$ is the temperature difference of the water flowing out of the HP and the water flowing into the HP. The former temperature is assumed to be 50°C and the latter temperature is retrieved from the sensors at $h_{\text{S,HP}}$ of the heat storage simulation. The water flow $F_{\text{S,D}}$ from the heat storage to the district is calculated via

$$F_{\text{S,D}} = \frac{D}{c_{\text{water}} \cdot \Delta_{\text{T,D}}} \tag{3}$$

where D is the current demand. $\Delta_{\text{T,D}}$ is the temperature difference of the water flowing into the district compared to the water flowing out of the district. For this value an assumption of 5°C is made. For conservation reasons it follows $F_{\text{S,HP}} = -F_{\text{HP,S}}$ for the water flow from the heat storage to the HP as well as $F_{\text{D,S}} = -F_{\text{S,D}}$ for the water flow from the district to the heat storage. The environment is idealized in such a way that the heat transfer is assumed to be loss free. Furthermore, there is no domestic hot water included in the heat demand of the district. The HP doesn't obey any locking times meaning that it can be freely operated by the agent. Figure 3 shows a schematic overview of the learning environment. The agent can choose its action from a continuous range from 0 kW to 100 kW. This is called the action space. To make its choice the agent sees an observation space that consists of multiple observables. Two of these observables are of endogenous nature since they are determined by the district heating model and the agent's action respectively. These two variables are the scalars SOC which denotes the state of charge of the heat storage and $loss_{\text{h}}$ which accounts for heat losses of the heat storage. The SOC stays in a range from 0 to 100%. The remaining observation space is of exogenous nature and consists of the scalars COP, the ambient temperature T_{amb} (see Figure 1) as well as the time features h, $minute$ and day. All of these scalars are depending on the current time step. The observation space also consists of two time series that provide the agent with information that is to be expected in the next 24 hours, thus 96 time steps. Firstly, the agent sees the future electricity prices $price_{\text{el}}$ obtained from the day-ahead market as explained in Section 2. Secondly, the agent sees the future demand D. In reality, the future demand can not be perfectly known. Therefore, over the course of this work, the following four cases are considered for D:

- **Perfect**: The next 96 values from the real data are taken

* **Persistence**: The previous 96 values from the data are taken as expected demand for the next 96 points in time
* **Forecast**: The forecasts as described in Section 3.6 are taken
* **No demand**: No demand is visible at all for the agent

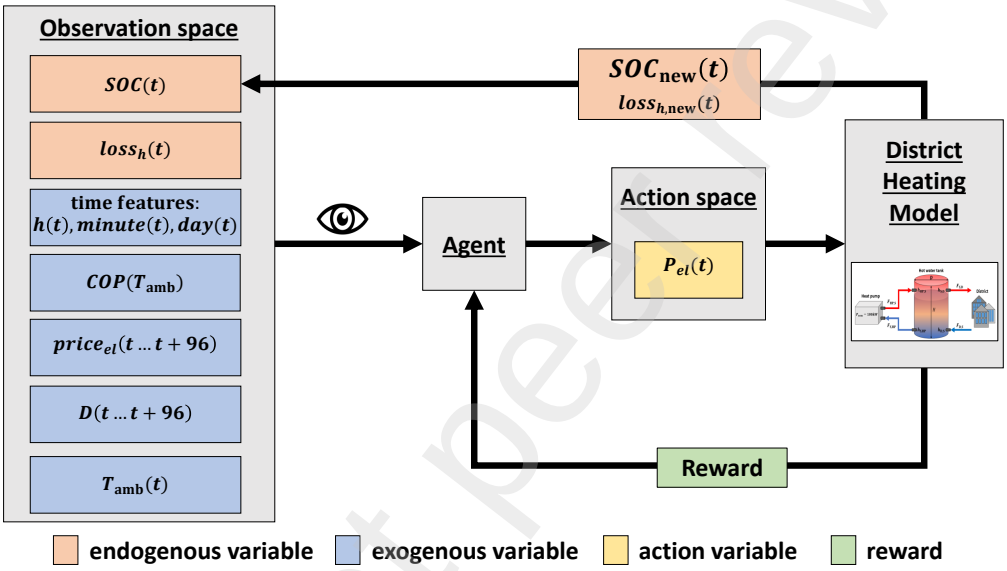

Figure 3: Schematic representation of the learning environment. Detailed picture of district heating model in Figure 2

Both the variables of the action space as well as the variables of the observation space are normalized to a range from -1 to 1.

3.5. Reward Function Design

The reward function consists of a positive part r_{pos} that rewards the agent to keep the SOC in a specific range as well as a negative part r_{neg} that penalizes the agent in form of energy costs. For the positive part of the reward function two different functions are used interchangeably for later comparisons:

Firstly, r_{pos} is described by a parabola with a maximum reward at an SOC value of 0.5 and a minimum reward at an SOC value of 0.01 and 0.99 (see Equation 4). For SOC values of less than 0.01 and more than 0.99 the

agent gets a negative reward of 10000 and is restarted. It is chosen so large to train the agent to never reach these low and high SOC values.

Secondly, r_{pos} is calculated with a step function with a step at a SOC value of 0.2. Below this threshold there is no reward while for SOC values larger than 0.2 there is a constant reward of 1 (see Equation 5).

$$r_{\text{pos}}^{\text{para}}(\text{SOC}) = -\frac{(\text{SOC} - 0.5)^2}{(0.01 - 0.5)^2} + 1 \tag{4}$$

$$r_{\text{pos}}^{\text{step}}(\text{SOC}) = \begin{cases} 0 & \text{if SOC} < 0.2 \\ 1 & \text{else} \end{cases} \tag{5}$$

A display of the two functions can be seen in Figure 4. The negative part of the reward function or penalty r_{neg} depends on the demand D, the heat loss $loss_{\text{h}}$ as well as the electricity price $price_{\text{el}}$ and can be expressed via

$$r_{\text{neg}} = \frac{-(D + loss_{\text{h}}) \cdot price_{\text{el}}}{\text{Euro}}. \tag{6}$$

The division by Euro is necessary to achieve a unitless reward function. The two final reward functions used in this study are thus

$$r_{\text{para}} = r_{\text{pos}}^{\text{para}} + r_{\text{neg}} \tag{7}$$

and

$$r_{\text{step}} = r_{\text{pos}}^{\text{step}} + r_{\text{neg}}. \tag{8}$$

As these two functions only differ in the positive part, they are referred to as *parabolic reward function* r_{para} and *step-shaped reward function* r_{step} in the following.

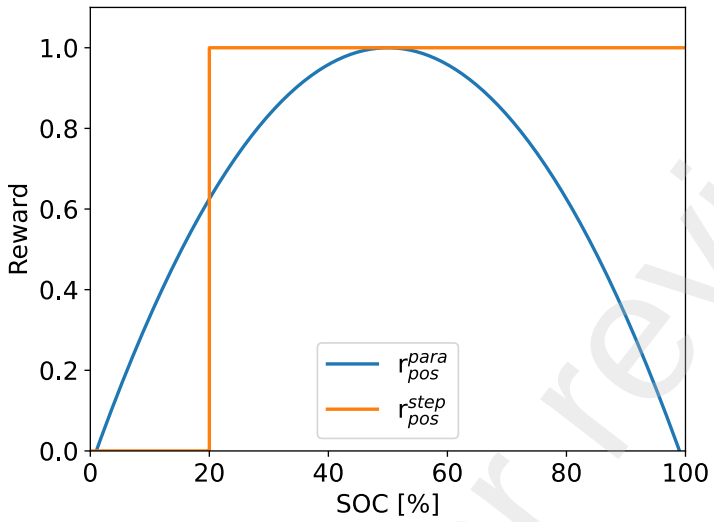

Figure 4: Representation of the positive part of the reward function.

3.6. Demand Forecast Creation

The demand forecast has been created using an LSTM trained on the first two of all five years of the heat demand data described in Section 2. The third year was used for validation of the LSTM while the fourth and fifth year are considered test set and can therefore be used as input for the RL agent. The hyperparameters of the LSTM have been found using optuna's TPESampler [36] performing 500 trials. The LSTM gets an input of 96 values which equals one day and outputs 96 as well to forecast the next day's demand. To assess the forecasts of the LSTM it is compared to the persistence forecasts which are 96 values of the previous day. The mean absolute error (MAE) is chosen as an evaluation metric. For every 96 values an MAE is calculated with the LSTM forecast and the persistence forecast. The average of all these values over the whole test set is then compared between both cases. The average MAE for the LSTM forecasts is 900.0 Wh and 998.9 Wh for the persistence forecasts. Thus, the LSTM forecast is about 10% better than the persistence forecast. From now on the LSTM forecast is referred to as *forecast* and the persistence forecast is referred to as *persistence*. The true data is denoted with the label *perfect*. Figure 5 shows the heat demand for a day in the winter and a day in the summer. It can be seen that on a winter day the *forecast* is generally slightly better than *persistence*. On a summer day, the *forecast* is worse since it

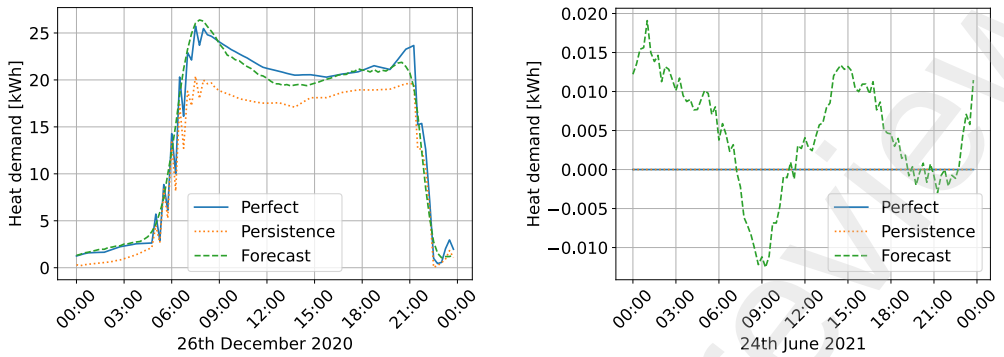

Figure 5: Heat demand in winter (left) and in summer (right).

fluctuates around zero. However, the prediction of space heating demands during summer is $< 1\%$ compared to demands during winter. Thus, these fluctuations in summer can be interpreted as noise of the prediction model. For the RL agent the negative values in such a case have been set to zero since a negative heat demand is not possible.

3.7. Benchmarking Against Rule-Based Operation

In order to benchmark the performance of the RL agent a rule-based approach to operation is used which will be referred to as *hysteresis* in the following. Hysteresis strategies are commonly applied in district heating systems and consist mainly of two rules or thresholds: A lower threshold of the SOC value of the heat storage where the HP starts to operate in order to increase the SOC as well as an upper threshold of the SOC value where the HP stops operating. In this work, the lower and upper thresholds are 20% and 100% respectively. When active for hysteresis operation, the HP is always operated at nominal power which is 100 kW in this work.

3.8. Evaluation Metrics

To evaluate the performance of the agent the following metrics have been chosen. They relate to energy costs as well as quantities concerning the HP and heat storage:

- $\mathbf{C}_{\text{tot}}$: Total energy costs
- $\mathbf{C}_{\text{con}}$: Energy costs due to energy consumption

* $\mathbf{C}_{\text{loss}}$: Energy costs due to heat loss
* $\mathbf{N}_{\text{on/off}}$: Number of on/off state changes of the HP
* $\mathbf{P}_{\text{avg}}$: Average operating power of the HP
* $\mathbf{P}_{\text{max}}$: Maximum operating power of the HP
* $\mathbf{SOC}_{\text{avg}}$: Average SOC of the hot water tank
* $\mathbf{SOC}_{\text{max}}$: Maximum SOC of the hot water tank

The most important measure is C_{tot} since it provides information about how cost efficient the agent is compared to the hysteresis operation. C_{tot} is the sum of C_{con} and C_{loss}. The latter two give insight about the distribution of costs. $N_{\text{on/off}}$ measures the amount of on/off state changes of the HP which is an important quantity to foresee its lifetime. A high number of on/off state changes can significantly reduce the lifetime. Finally, the *mean* and *max* of the HP's power as well as the heat storage's SOC are looked at. These values provide information about their sizing.

3.9. Learning Setup

The PPO algorithm as described in Section 3.1 has mainly been used with its default settings which can be found in this documentation [37]. However, the following hyperparameters were adapted since a smoother learning curve and a faster learning has been observed when using them.

- The learning rate is set to decrease from 0.001 to 0 along the learning process.
- The parameter n_{steps} describes after how many observed steps the policy is being updated and has been set to one year of data. The default value of this parameter is 2048 which corresponds to roughly 21 days of the given data set at 15 min granularity. In our case, this leads to fluctuations in the learning process. This is mostly likely due to the fact that in 21 days the agent does not see enough variations of the heat demand compared to what will occur over the whole year.
- The batch size has been chosen to be 10 days (960 time steps) which speeds up the learning process compared to the default batch size of 64 which in our case equals only 18 hours.

All other hyperparameters are kept at their default values as suggested by Stable-Baselines3 [32]. To further speed up the training process the environment has been vectorized to train on 10 environments in parallel. The last year (2021) of the given five years of data serves as a test set and will not be seen by the agent during training. In order to save the best model a callback that frequently checks the model performance on the test set is used. Finally, it has to be addressed that PPO is strongly depending on the random seed. As a result, the learned policies can differ heavily among different training runs with different seeds. Therefore, ten agents have been trained independently from each other for each of the four cases *perfect*, *persistence*, *forecast* and *no demand*.

4. Results

This section firstly shows the results of the RL agents with regard to the evaluation metrics defined in Section 3.8. All shown metrics are based on the mean and standard deviation of the individual agents in order to show the robustness of the algorithms. Subsequently, we will have a look at how the different agents operate and which strategies have been learned. To produce these results the trained agents have been tested on the before unseen test set as described in Section 3.9.

4.1. Evaluation of RL agent

The total energy costs C_{tot} of the RL agents as well as of the hysteresis operation on the test set for both reward functions can be seen in Figure 6. The ratio between costs due to consumption and costs due to heat loss is visualized for every case. Additionally, the number of on/off state changes of the HP $N_{\text{on/off}}$ can be seen. The cases *perfect*, *persistence* and *forecast* result in approximately 10000 Euro of total energy costs with no significant differences within the error bars. Generally, the mean total energy costs for the parabolic reward function are slightly lower than the ones using the step-shaped reward function. The costs due to heat losses are approximately 2000 Euro for the parabolic reward function and 1500 Euro for the step-shaped reward function, respectively. The total energy costs for the case *no demand* amounts to approximately 11000 Euro in case of the parabolic reward function and 12000 Euro in case of the step-shaped reward function. The hysteresis operation causes total energy costs of almost 15000 Euro. The number of on/off state changes of the HP in the cases *perfect*, *persistence*,

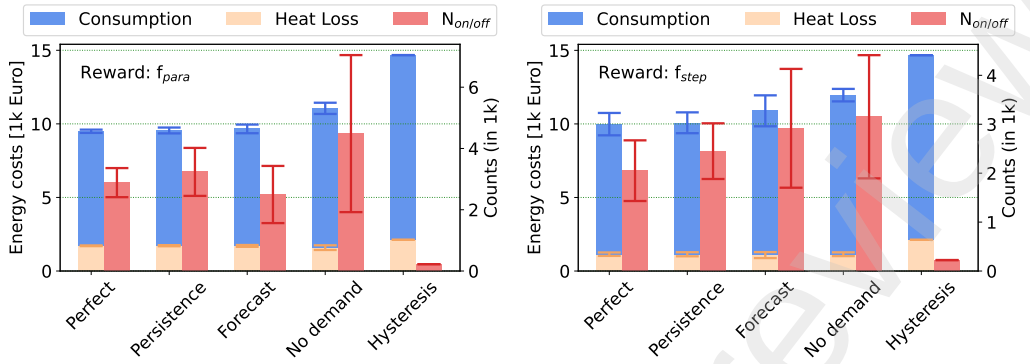

Figure 6: Total energy costs of the optimized agent using a parabola as the reward function (left) and a step function as the reward function (right) for the four different environments and the hysteresis operation. The total energy costs are divided in costs due to consumption (blue) and costs due to heat loss (yellow).

forecast and *no demand* fluctuates between values of 1000 and 7000 while the hysteresis operation only causes 224 on/off switches. The high error of these values is caused by the different policies the agent has learned. A closer look at this behavior can be seen in Figure 7 which shows the operation of the HP and heat storage of two different policies over one day. While $N_{\mathrm{on/off}}$ is different by almost a factor of four C_{tot} has about the same value. A correlation between $N_{\mathrm{on/off}}$ and C_{tot} could not be observed.

The remaining metrics are shown in Table 2. It can be seen that the average HP operating power of all RL agents is below 10 kW. By definition the hysteresis control always operates at 100 kW. The maximum power ever used by the RL agents is in a range of about 23 to 41 kW with respect to the different cases. The average SOC is around 45 % for the parabolic reward function and around 30 % for the step-shaped reward function which explains the smaller ratio of heat losses in the latter case. The average SOC of the hysteresis operation is at 55 % which causes the higher ratio of heat losses in the total energy costs. The maximum value of the SOC ever reached differs among all cases and lies in the range of about 58 to 94 %.

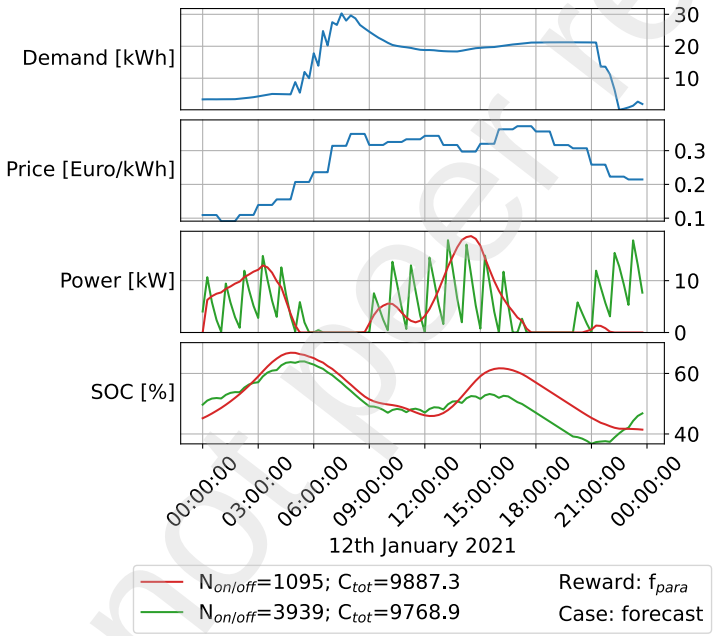

Figure 7: Operation of two different policies for the case *forecast* with the parabolic reward function.

	$\mathbf{P}_{\text{avg}}[kW]$	$\mathbf{P}_{\text{max}}[kW]$	$\mathbf{SOC}_{\text{avg}}[\%]$	$\mathbf{SOC}_{\text{max}}[\%]$
Hysteresis	100.0(0.0)	100.0(0.0)	55.0(0.0)	100.0(0.0)
	Reward function: parabola			
Perfect	9.5(1.2)	36.2(5.1)	46.1(0.9)	83.1(4.4)
Persistence	9.9(1.4)	41.1(9.4)	46.0(0.8)	87.3(6.5)
Forecast	8.3(1.4)	28.7(4.3)	45.8(1.0)	88.2(8.7)
No demand	7.7(2.5)	39.8(18.5)	41.5(4.5)	69.2(6.9)
	Reward function: step function			
Perfect	9.0(2.6)	30.2(8.4)	30.9(3.4)	89.5(13.8)
Persistence	9.6(1.5)	34.4(7.4)	30.4(3.4)	94.0(4.7)
Forecast	7.3(1.4)	24.2(3.9)	29.4(6.5)	80.0(16.8)
No demand	5.3(0.4)	22.8(5.4)	28.4(3.6)	58.1(10.7)

Table 2: Mean and standard deviation of the evaluation metrics described in Section 3.8.

4.2. Operation of RL agent

The results shown in this section are based on the respective run with the lowest total energy costs. Figure 8 shows the agents actions on the HP as well as the behavior of the heat storage for a week with a high heat demand. The same analysis for a week with a low heat demand can be seen in Figure 9. In both plots the results on the left hand side have been produced with the parabolic reward function while the right hand side uses a step-shaped reward function.

It can clearly be seen that the agent learned to avoid operating the HP when the electricity price is high. The charts also display the behavior of the SOC that is intended by the reward function. For the parabolic reward function case, the agent tries to keep the SOC at around 50% while for the step-shaped reward function case the agent tries to keep the SOC just above 20%. The behavior of the SOC among the three cases *forecast*, *persistence* and *perfect* is very similar while for the case *no demand* the agent reflects a slightly different behavior. Nevertheless, even if the RL agent does not see any demand information, it still finds an operational strategy with significantly lower energy costs compared to the hysteresis operation. It can also be observed in Figure 9 that the power is fluctuating a lot. The reason for this is the design of the reward function. At time steps the HP is operated the agent is penalized according to Equation 6. To reach a positive reward for its action the agent tries to select the action so small that the reward due to the SOC value as described by Equation 4 and Equation 5 is not exceeded.

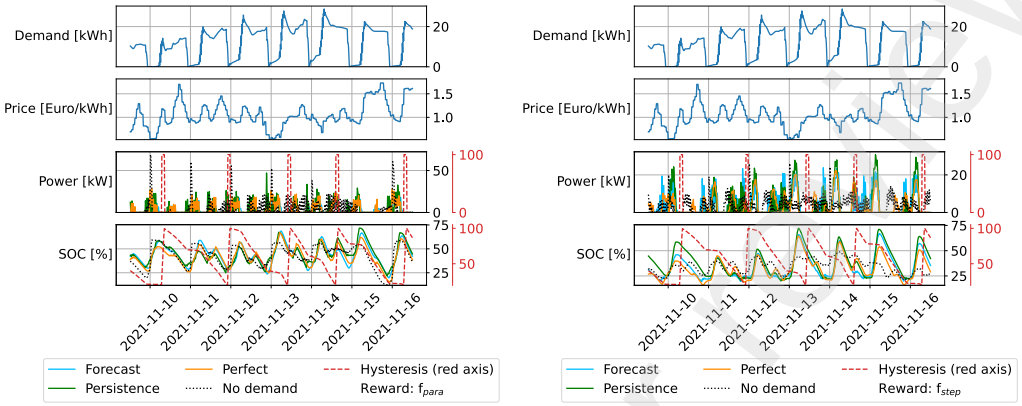

Figure 8: Operation of the HP with the optimized agent on a week with a high heat demand using a parabola as the reward function (left) and a step function as the reward function (right) for the four different cases and the hysteresis operation.

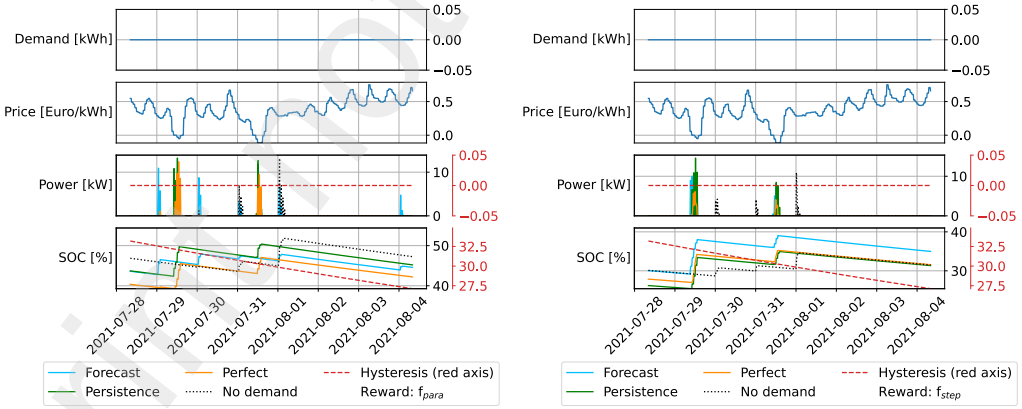

Figure 9: Operation of the HP with the optimized agent on a week with no heat demand using a parabola as the reward function (left) and a step function as the reward function (right) for the four different cases and the hysteresis operation.

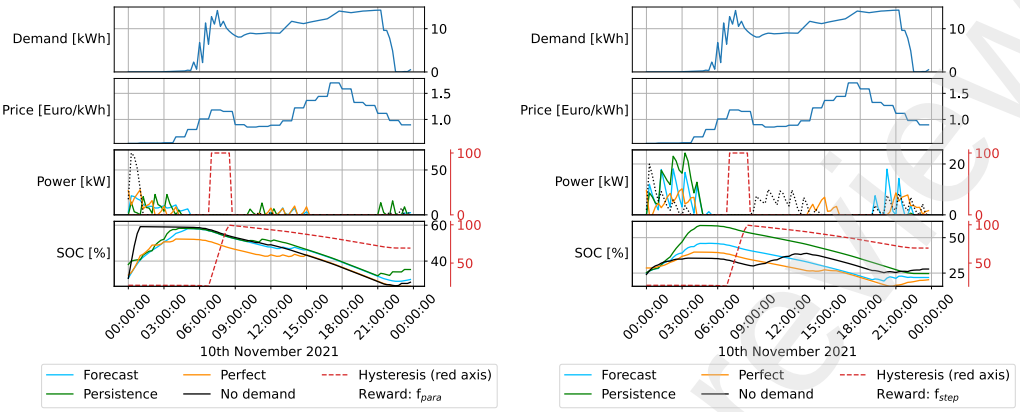

Figure 10: Operation of the HP with the optimized agent on a winter day using a parabola as the reward function (left) and a step function as the reward function (right) for the four different cases and the hysteresis operation.

A closer look at these fluctuations can be seen in Figure 10.

5. Discussion

The RL agent has learned to exploit the variable electricity prices very well during summer but also during winter. This leads to significantly lower energy costs compared to a conventional hysteresis operation. The energy savings of the RL agents using the parabolic reward function for the case *persistence* compared to the hysteresis operation reach 34.9 %. A similar work using rainbow deep reinforcement learning reports a reduction of 22.2 % of electricity costs compared to a rule-based control [38]. In another work reinforcement learning is used to control a HVAC system with a regular thermostat control with the findings of an approximate cost reduction of 15 % comparing the two approaches [39].

Reviewing our approach we find that the parabolic reward function slightly outperforms the step-shaped reward function by means of total energy costs. Additionally the parabolic reward function effects that the average SOC is between 40 % and 50 % ensuring a greater flexibility of the storage.

As a mandatory condition of the environment, demand is met at all times, therefore living comfort is never compromised. Comparable work often requires a building model like [40] whereas our approach only relies on the future demands, electricity prices and temperature information. All of the

above are easily attainable by e.g., using simple persistence forecasts, available day-ahead prices and publicly available weather forecasts. Thus, our approach would likely increase the acceptability of RL as operational management technique in real world applications.

A very interesting observation is that the agent performs equally well in the cases *perfect*, *persistence* and *forecast*. When forecasts are available, the RL agent decreases operating electricity costs by up to approximately 35% with differences being insignificant regarding the three different forecast cases. Unlike other publications like [41] and [42] that stress the importance of accurate forecasts our approach does not require high quality forecasts but only a rough prognosis like the *persistence* case. Even if there is no forecast at all the agent performs a lot better than the hysteresis operation: For the parabolic reward function energy savings result to 24.7% and for the step-shaped reward functions the energy savings are 18.5%. This is due to the installed hot water tank which is big enough to provide enough inertia and flexibility to compensate for missing or slightly incorrect demand information. However, it is likely that the importance of quality of demand forecasts increases at smaller storage capacities.

Looking at the maximum power used by the agent it gets clear that the installed nominal power of the district's HP is not fully utilized. Dependent on the RL agent's setup, only $23-41\%$ of the heat pump's installed nominal power of 100 kW are exploited. With an intelligent operational management as we show in this work, the size of the installed HP could be therefore potentially decreased as long as security concerns for peak demands are still taken into account. Note, that decreasing the installed nominal power would only affect the hysteresis operation in the first place as long as the nominal power is greater than the maximum power used by the RL agents.

Nevertheless, the hysteresis approach requires a lot less state switches of the HP compared to the operation of all RL agents. Frequent state switches increase degradation of HP components and would lead to higher maintenance costs. Therefore, the reward function could eventually be adapted to penalize frequent state switches more to ensure a smoother operation.

The environment used for the presented results has been idealized. The heat transfer between the HP/district and the hot water tank has been assumed to be loss free. While for the purpose of this work it is a valid assumption one should incorporate this loss for future work. If this can not be addressed by a simulation one could simply choose a constant value to represent this loss for each time step. Additionally, we assumed the temperature

difference between the water flowing into the district and out of the district to always be 5 °C in order to determine the SOC of the heat storage. This is realistic since the temperature spread can be fixed in a real world building by setting the flow velocity of the warm water through the heating pipes inside the building which is called hydraulic levelization.

We also didn't incorporate locking periods of the HP after a state switch which could be implemented in future work. Since in Germany specifically energy suppliers can reserve the right to disconnect the HP from the grid for the sake of grid stability, this external steering signal could be taken into account in further studies.

Lastly, the presented approach does not require any building information other than the respective demand data and very basic measurements of the installed hot water tank. In order to apply our approach into real world systems, one could firstly collect demand data for a certain period of time - ideally for at least one year to get data from all seasons - while still operating the HP using a classical rule-based approach. In parallel, the RL agent could be pre-trained and take over at a certain point in time. Since with our approach we also observed significant energy savings without demand forecasts one could even reject the rule-based approach and use the RL agent immediately. The required IT infrastructure for this could easily be installed on-site near the HP. As a safety measure, the rule-based approach could always serve as a fallback solution that kicks in when specific parameters are met. The training of the agent was performed on a *NVIDIA Quadro RTX 6000* and took around 24 hours for one year of data where the heat storage simulation is the main bottleneck. Therefore, the training would have to be performed off-site or cloud-based while in production the agent is fast enough to work on-site. At least in production, the demand data can be processed decentrally on-site near the HP operation. Security and safety concerns are thereby minimized.

6. Conclusion and Outlook

This work shows a successful utilization of a RL approach to operate a HP in a residential district. It has been discovered that such an approach can significantly reduce energy costs by approximately 35 %. Additionally, we show that the intelligent operation of HPs does not use the full installed nominal power and could therefore reduce investment costs. We investigated the impact of demand forecasts on the results of a RL-based operation of

the respective HP and find that the quality of demand forecasts is only of minor importance. Even agents having no demand information at all still exceed a rule-based approach significantly. Two different reward functions are applied. A parabolic reward function leads to a RL-based operation of the HP keeping the SOC of the heat storage at around 50 % which could enable further business models of selling upward and downward flexibility to the grid operators. The RL agent and its reward function respectively could also be expanded to account for this business model. On the other hand, a step-shaped reward function leads to a RL-based operation that uses the full flexibility of the heat storage to minimize energy costs especially due to losses in the heat storage. The high robustness and repeatability of results is proven by showing means and standard deviations of all evaluation metrics based on ten independent runs of the RL agents. Although the learned policies differ significantly in their number of state changes of the HP, energy costs are very similar for each run.

Improvements for further studies could be to increase the complexity of the environment. In Section 3.4 a few idealizations have been mentioned that could be replaced with more sophisticated information. One example would be to include heat losses during heat transfer. Another one is to take into account locking periods in which the state of the HP cannot be changed after a switch occurred. A discretization of the agent's action space would additionally enable other algorithms than PPO to be applied to the given control problem.

Besides a business model to sell flexibility to the grid operators, maximizing the own consumption of a given PV system could be possible by widening the action space of the RL agent. Also multiple HPs and/or multiple heat storages can be considered.

Furthermore, the results of this work are based on space heating data only. Thus, it would be interesting to see the performance when domestic hot water is included. This would not change the complexity of the control problem but would only change the given demand time series to be more erratic. We expect that in this case the maximum power needed for the HP will roughly double.

References

[1] Federal Environment Agency Germany (Umweltbundesamt), Energy consumption by energy source and sec-

tor, availabe at: https://www.umweltbundesamt.de/daten/energie/energieverbrauch-nach-energietraegern-sektoren#allgemeine-entwicklung-und-einflussfaktoren (2021).

[2] Federal Environment Agency Germany (Umweltbundesamt), Energy consumption of private households, availabe at: https://www.umweltbundesamt.de/daten/private-haushalte-konsum/wohnen/energieverbrauch-privater-haushalte#hochster-anteil-am-energieverbrauch-zum-heizen (2021).

[3] J. Deason, M. Borgeson, Electrification of buildings: potential, challenges, and outlook, Current Sustainable/Renewable Energy Reports 6 (2019) 131–139.

[4] H. Kazmi, S. D'Oca, Demonstrating model-based reinforcement learning for energy efficiency and demand response using hot water vessels in net-zero energy buildings, in: 2016 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), IEEE, 2016, pp. 1–6.

[5] G. Mbiydzenyuy, S. Nowaczyk, H. Knutsson, D. Vanhoudt, J. Brage, E. Calikus, Opportunities for machine learning in district heating, Applied Sciences 11 (13) (2021) 6112.

[6] S. Noye, R. M. Martinez, L. Carnieletto, M. De Carli, A. C. Aguirre, A review of advanced ground source heat pump control: Artificial intelligence for autonomous and adaptive control, Renewable and Sustainable Energy Reviews 153 (2022) 111685.

[7] C. Ntakolia, A. Anagnostis, S. Moustakidis, N. Karcanias, Machine learning applied on the district heating and cooling sector: A review, Energy Systems (2021) 1–30.

[8] Z. Wang, T. Hong, Reinforcement learning for building controls: The opportunities and challenges, Applied Energy 269 (2020) 115036.

[9] L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A survey, Journal of artificial intelligence research 4 (1996) 237–285.

[10] K. Mason, S. Grijalva, A review of reinforcement learning for autonomous building energy management, Computers & Electrical Engineering 78 (2019) 300–312.

[11] T. Peirelinck, F. Ruelens, G. Decnoninck, Using reinforcement learning for optimizing heat pump control in a building model in modelica, in: 2018 IEEE International Energy Conference (ENERGYCON), IEEE, 2018, pp. 1–6.

[12] X. Yuan, Y. Pan, J. Yang, W. Wang, Z. Huang, Study on the application of reinforcement learning in the operation optimization of hvac system, in: Building Simulation, Vol. 14, Springer, 2021, pp. 75–87.

[13] G. Pinto, M. S. Piscitelli, J. R. Vázquez-Canteli, Z. Nagy, A. Capozzoli, Coordinated energy management for a cluster of buildings through deep reinforcement learning, Energy 229 (2021) 120725.

[14] C. Correa-Jullian, E. L. Droguett, J. M. Cardemil, Operation scheduling in a solar thermal system: A reinforcement learning-based framework, Applied energy 268 (2020) 114943.

[15] P. Lissa, M. Schukat, M. Keane, E. Barrett, Transfer learning applied to drl-based heat pump control to leverage microgrid energy efficiency, Smart Energy 3 (2021) 100044.

[16] F. Ruelens, S. Iacovella, B. J. Claessens, R. Belmans, Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning, Energies 8 (8) (2015) 8300–8318.

[17] B. V. Mbuwir, D. Geysen, F. Spiessens, G. Deconinck, Reinforcement learning for control of flexibility providers in a residential microgrid, IET Smart Grid 3 (1) (2020) 98–107.

[18] A. Heidari, F. Marechal, D. Khovalyg, An adaptive control framework based on reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems, in: Journal of physics: Conference series, Vol. 2042, IOP Publishing, 2021, p. 012006.

[19] D. Pujić, M. Jelić, M. Batić, N. Tomašević, Application of reinforcement learning for control of heat pump systems (2022).

[20] L. Langer, T. Volling, A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems, Applied Energy 327 (2022) 120020.

[21] S. Hummel, C. Betzold, A. Dentel, Impact of the weather forecast quality on a mpcdriven heat pump heating system, in: CLIMA 2022 conference, 2022.

[22] A. Dreher, T. Bexten, T. Sieker, M. Lehna, J. Schütt, C. Scholz, M. Wirsum, Ai agents envisioning the future: Forecast-based operation of renewable energy storage systems using hydrogen with deep reinforcement learning, Energy Conversion and Management 258 (2022) 115401.

[23] Generische Gebäudesimulation als Bestandteil der Quartier-Simulationssoftware "QuaSi"-Verbundvorhaben EnStadtEs-West: Klimaneutrales Stadtquartier Neue Weststadt Esslingen (2020).

[24] DIN e.V., DIN 4108-6, Wärmeschutz und Energie-Einsparung in Gebäuden - Teil 6: Berechnung des Jahresheizwärme- und des Jahresheizenergiebedarfs (2003).

[25] D. Crawley, C. Pedersen, L. Lawrie, F. Winkelmann, Energyplus: Energy simulation program, ASHRAE Journal 42 (2000) 49–56.

[26] [dataset] - Open Data Platform of Deutscher Wetterdienst, available at: https://www.dwd.de/DE/leistungen/opendata/opendata.html.

[27] SMARD, Marktdaten (2023).
URL https://www.smard.de

[28] Federal Network Agency Germany (Bundesnetzagentur) (2023). [link].
URL https://www.bundesnetzagentur.de

[29] ENTSO-E, European network of transmission system operators for electricity (2023).
URL https://www.entsoe.eu

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms (2017). arXiv:1707.06347.

[31] C. J. C. H. Watkins, P. Dayan, Q-learning, Machine Learning 8 (3) (1992) 279–292. doi:10.1007/BF00992698.

[32] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann, Stable-baselines3: Reliable reinforcement learning implementations, Journal of Machine Learning Research 22 (268) (2021) 1–8.

[33] K. S. J. Gerster, M. Blank, M. Sonnenschein, Intelligentes heimenergiemanagement – nutzung der synergiepotentiale bei der thermischen und elektrischen objektversorgung durch modellbasierte und prädiktive betriebsführungsstrategien, 2016.

[34] P. Kasturi, J. S. Schwarz, mosaik-heatpump, https://gitlab.com/mosaik/components/energy/mosaik-heatpump.

[35] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, Openai gym (2016). arXiv:arXiv:1606.01540.

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[37] S.-B. Contributors, Stable baselines3 documentation release 1.5.0 (2023).
URL https://stable-baselines3.readthedocs.io/_/downloads/en/v1.5.0/pdf/

[38] G. Han, H.-J. Joo, H.-W. Lim, Y.-S. An, W.-J. Lee, K.-H. Lee, Data-driven heat pump operation strategy using rainbow deep reinforcement learning for significant reduction of electricity cost, Energy 270 (2023) 126913. doi:https://doi.org/10.1016/j.energy.2023.126913.

[39] T. Peirelinck, F. Ruelens, G. Decnoninck, Using reinforcement learning for optimizing heat pump control in a building model in modelica, in: 2018 IEEE International Energy Conference (ENERGYCON), 2018, pp. 1–6. doi:10.1109/ENERGYCON.2018.8398832.

[40] L. Yang, Z. Nagy, P. Goffin, A. Schlueter, Reinforcement learning for optimal control of low exergy buildings, Applied Energy 156 (2015) 577–586. doi:https://doi.org/10.1016/j.apenergy.2015.07.050.

[41] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, J. Liu, Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms, Energy 188 (2019) 116085. doi:https://doi.org/10.1016/j.energy.2019.116085.

[42] F. Bünning, P. Heer, R. S. Smith, J. Lygeros, Improved day ahead heating demand forecasting by online correction methods, Energy and Buildings 211 (2020) 109821. doi:https://doi.org/10.1016/j.enbuild.2020.109821.

CRediT authorship contribution statement

Simon Schmitz: Writing – original draft, Writing – review & editing, Conceptualization, Formal Analysis, Methodology, Investigation, Visualization, Data curation, Software **Karoline Brucke:** Project administration, Conceptualization, Writing – original draft, Investigation, Validation, Formal Analysis. **Pranay Kasturi:** Software, Writing – original draft. **Esmail Ansari:** Data curation, Writing – original draft. **Peter Klement:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement and funding

This work is supported by the Helmholtz Association's Initiative and Networking Fund (INF) under the Helmholtz AI platform grant agreement (ID ZT-I-PF-5-1), Local Unit 'Munich Unit @Aeronautics, Space and Transport (MASTr)' as well as the German Federal Ministry for Economic Affairs and Climate Action (BMWK) and the Federal Ministry of Education and Research (BMBF) in the project ENaQ (project number 03SBE111).