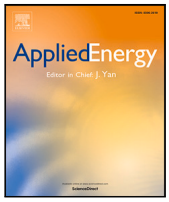
Applied Energy 371 (2024) 123688

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Forecast-based and data-driven 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

ARTICLE INFO

Keywords:
Reinforcement learning
Heat pump operation
Residential heating
Demand forecast
Operation under uncertainty

ABSTRACT

Electrified residential heating systems have great potential for flexibility provision to the electricity grid with advanced operation control mechanisms being able to harness that. In this work, we therefore apply a reinforcement learning (RL) approach for the operation of a residential heat pump in a simulation study and compare the results with a classical rule-based approach. Doing so, we consider an apartment complex with 100 living units and a central heat pump along with a central hot water tank serving as heat storage. Unlike other studies in the field, we focus on a data driven approach where no building model is required and living comfort of the residents is never compromised. Both factors maximize the applicability in real world buildings. Additionally, we examine the effects of uncertainty on the heat pump operation. This is carried out by testing four different observation spaces each with different data visibility and availability to the RL agent. With that we also simulate the heat pump operation under forecast conditions which has not been done before to the best of our knowledge. We find that the inertia of typical residential heat systems is high enough so that missing or uncertain information has only a minor effect on the operation. Compared to the rule-based approach all RL agents are able to exploit variable electricity prices and the flexibility of the heat storage in such a way, that electricity costs and energy consumption can be significantly reduced. Additionally, a large proportion of the nominal electrical power of the installed heat pump could be saved with the presented intelligent operation. The robustness of the approach is shown by running ten independent training and testing cycles for all setups with reproducible results.

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 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 (HWT) 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.

* Corresponding author.
E-mail address: Simon.Schmitz@dlr.de (S. Schmitz).

https://doi.org/10.1016/j.apenergy.2024.123688
Received 2 February 2024; Received in revised form 22 May 2024; Accepted 9 June 2024

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 HWTs 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 [11,16,17]. 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 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 HP for space heating in a big apartment complex. 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 HWT 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 HWT 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 HWT. No building envelope and therefore also no indoor temperature is modeled. This approach 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 examining 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 HWT 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 apartment complex's electricity 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 apartment complex 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

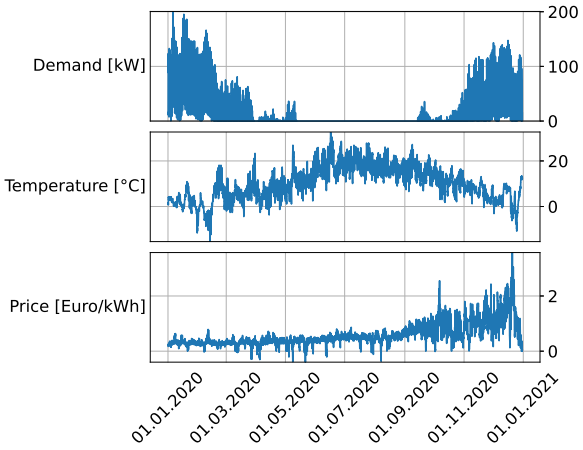


Fig. 1. Heat demand, ambient temperature and electricity prices for a one year period. (Time interval: 15 min)

01.01.2020 to 01.01.2021 can be seen in Fig. 1. As can be seen the heat demand reaches values of up to 50 kWh per 15 min 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.

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. Fig. 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. Note, that the electricity prices obtained from SMARD do not include grid fees and other charges.

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] as shown in Fig. 2. Stable-Baselines3 is a Python package providing implementations of multiple reinforcement learning algorithms. In our work, it is used for modeling and training

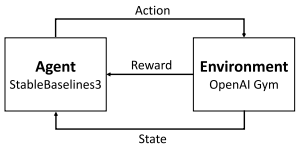


Fig. 2. Conceptual representation of the interplay of the used reinforcement learning python libraries.

the agent and its embedding into the learning environment. The OpenAI Gym interface on the other hand is used for modeling the environment itself [33]. This includes the definition of the observation space and the agent's action space.

3.2. Hot water tank model

The HWT model used for this work was developed for another project [34] and is openly available as a part of the mosaik-heatpump repository [35]. 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.

The heat storage for the considered apartment complex is modeled by one central HWT and can be seen in Fig. 3. It has a height of $H = 5\,\text{m}$ and a diameter of $D = 4\,\text{m}$. These dimensions result in a volume of almost 63 000 l. The HWT has connectors at $h_{\text{HP,S}} = h_{\text{S,D}} = 4.999\,\text{m}$ for the hot water and connectors at $h_{\text{S,HP}} = h_{\text{D,S}} = 0.001\,\text{m}$ for the cool water. Three layers are considered for modeling the stratification and the ambient temperature of the HWT is set to be at constant 20 °C.

3.3. Heat pump model

The HP is of type air-to-water and 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 COP values 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 electrical operational power P_{el} of the HP as follows.

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

P_{th} is used to determine the water flow $F_{\text{HP,S}}$ from the HP to the HWT (see Fig. 3). For our work, P_{el} can be continuously set in a range from 0 to 100 kW.

Table 1
COP values for specific ambient temperatures T_{amb} and outflow water temperatures T_w of the HP taken from the manufacturer's design tool [36].

COP values	T_{amb} [°C]						
T_w [°C]	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	–

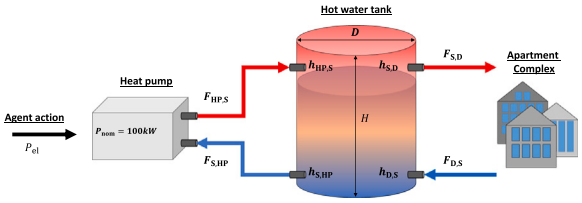


Fig. 3. Schematic representation of the used heating model including a HWT as heat storage.

The installed nominal electrical HP power results from the following: The apartment complex 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 electrical 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 electrical power are assumed here.

3.4. Environment design

The environment described in this section has been built with Stable-Baselines3 [32] in combination with OpenAI's Gym library [33]. The interplay of the two libraries is depicted in Fig. 2 and explained in Section 3.1. The environment consists of a HP simulation as described in Section 3.3 as well as a HWT simulation as described in Section 3.2. A schematic overview of the simplified heat network of the apartment complex can be seen in Fig. 3. The assumptions regarding the heat network of the apartment complex are based on the following publications by Klement et al. [37] and Schmeling et al. [38].

The water flow $F_{HP,S}$ between the HP and the HWT is calculated by

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

where c_{water} denotes the specific heat capacity of water and $\Delta_{T,HP}$ 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 $h_{S,HP}$ of the HWT simulation. The water flow $F_{S,D}$ from the HWT to the apartment complex is calculated via

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

where D is the current demand. $\Delta_{T,D}$ is the temperature difference of the water flowing into the heat exchangers of the apartment complex compared to the water flowing out of the heat exchangers of the apartment complex. This temperature spread is determined by the heat exchangers of the heating system in the apartment complex and is fixed to 5 °C. For conservation reasons it follows $F_{S,HP} = -F_{HP,S}$ for the water flow from the HWT to the HP as well as $F_{D,S} = -F_{S,D}$ for the water flow from the apartment complex to the HWT. Note, that the heating system of the apartment complex is not directly connected to the flow out of the HWT since thermal and hydraulic decoupling by heat exchangers takes place. The environment is idealized in such a way that the heat transfer between HP and HWT is assumed to be loss free. Losses within the apartment complex and due to the thermal decoupling at the heat exchangers are included in the demand data. Furthermore, there is no domestic hot water included in the heat demand of the apartment complex. The HP does not obey any locking times meaning that it can be freely operated by the agent. Fig. 4 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 heating model and the agent's action respectively. These two variables are the scalars SOC which denotes the state of charge of the HWT and $loss_h$ which accounts for heat losses of the HWT. The SOC stays in a range from 0 to 100% and is determined via the mean temperature of the HWT T_{mean} according to the following equation.

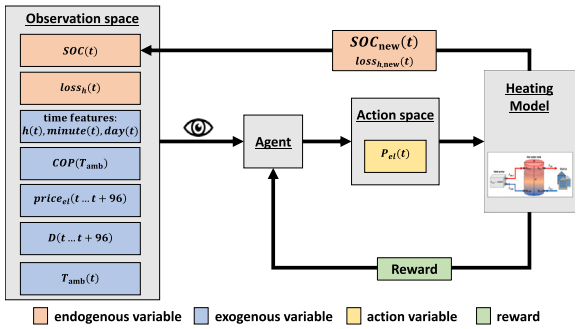


Fig. 4. Schematic representation of the learning environment. A detailed picture of the heating model is shown in Fig. 3.

$$SOC = \frac{T_{mean} - T_{min}}{T_{max} - T_{min}} \tag{4}$$

Thus, if $T_{mean} = T_{min}$ the SOC resolves to 0 % while in case of $T_{mean} = T_{max}$ the SOC resolves to 100 %. T_{min} is chosen to be 20 °C which is the ambient temperature of the HWT. T_{max} is chosen to be 50 °C which is the water temperature that is provided by the HP. The remaining observation space is of exogenous nature and consists of the scalars COP, the ambient temperature T_{amb} (see Fig. 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 h, thus 96 time steps. Firstly, the agent sees the future electricity prices $price_{el}$ obtained from the day-ahead market as explained in Section 2. Secondly, the agent sees the future demand D. In reality, the future demand cannot 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

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 electricity costs. For the positive part

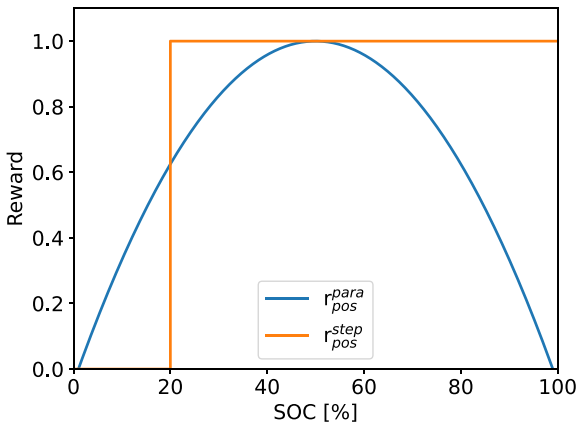


Fig. 5. Representation of the positive part of the reward function.

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 Eq. (5)). 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 Eq. (6)).

$$r_{\text{pos}}^{\text{para}}(\text{SOC}) = -\frac{(\text{SOC}-0.5)^2}{(0.01-0.5)^2} + 1 \tag{5}$$

$$r_{\text{pos}}^{\text{step}}(\text{SOC}) = \begin{cases} 0 & \text{if SOC} < 0.2 \\ 1 & \text{else} \end{cases} \tag{6}$$

A display of the two functions can be seen in Fig. 5. 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

$$r_{\text{neg}} = \frac{-(D + loss_{\text{h}}) \cdot price_{\text{el}}}{\text{Euro}}. \tag{7}$$

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

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

and

$$r_{\text{step}} = r_{\text{pos}}^{\text{step}} + r_{\text{neg}}. \tag{9}$$

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.

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 [39] performing 500 trials. The LSTM gets an input of 96 values which equals one day and outputs 96 values 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*. Fig. 6 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, due to the absence of domestic hot water, no heat demand is expected. The fluctuations of the *forecast* in the right hand plot of Fig. 6 are caused by the LSTM that struggles to predict exactly zero. As can be seen from the y-axis scale the *forecast* 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 residential heating systems and consist mainly of two rules or thresholds: A lower threshold of the SOC value of the HWT 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 electrical 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 electricity costs as well as quantities concerning the HP and HWT:

- $\mathbf{C}_{\text{tot}}$: Total electricity costs
- $\mathbf{C}_{\text{con}}$: Electricity costs due to heat demand
- $\mathbf{C}_{\text{loss}}$: Electricity costs due to heat loss
- $\mathbf{E}_{\text{tot}}$: Total electricity consumption of the HP
- $\mathbf{N}_{\text{on/off}}$: Number of on/off state changes of the HP
- $\mathbf{P}_{\text{avg}}$: Average operating electrical power of the HP
- $\mathbf{P}_{\text{max}}$: Maximum operating electrical power of the HP
- $\mathbf{SOC}_{\text{avg}}$: Average SOC of the HWT
- $\mathbf{SOC}_{\text{max}}$: Maximum SOC of the HWT

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. Another important measure is E_{tot} because it can be used to assess how much energy can be saved with an intelligent control compared to a rule-based approach. Since the demand of the apartment complex is fixed E_{tot} enables an evaluation of the heat losses of the HWT. With that it can be assessed how much of the electricity cost savings are due to exploitation of electricity prices and how much are due to energy savings. $N_{\text{on/off}}$ 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 electrical power as well as the HWT's SOC are looked at. These values provide information about their sizing.

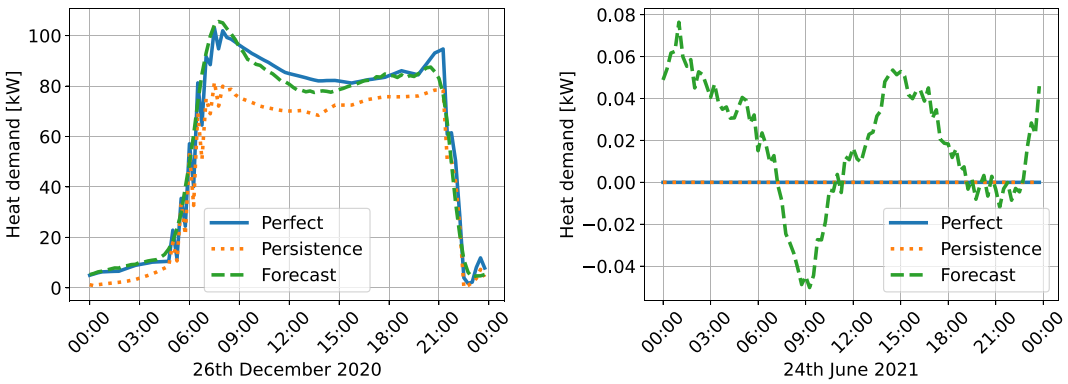


Fig. 6. Heat demand in winter (left) and in summer (right). (Time interval: 15 min)

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 [40]. 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 h.

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 electricity 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 Fig. 7. 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 electricity costs with no significant differences within the error bars. Generally, the mean total electricity 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 electricity costs for the case *no demand* amounts to approximately 11000 Euro in case of the parabolic reward function and 12,000 Euro in case of the step-shaped reward function. The hysteresis operation causes total electricity costs of almost 15,000 Euro. The number of on/off state changes of the HP in the cases *perfect*, *persistence*, *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 Fig. 8 which shows the operation of the HP and HWT of two different policies over one day. While $N_{\text{on/off}}$ is different by almost a factor of four C_{tot} has about the same value. A correlation between $N_{\text{on/off}}$ and C_{tot} could not be observed.

The remaining metrics are shown in Table 2. It can be seen that the average HP operating electrical power of all RL agents is below 10 kW. By definition the hysteresis control always operates at 100 kW. The maximum electrical 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 costs due to heat losses in the total electricity costs. The maximum value of the SOC ever reached differs among all cases and lies in the range of about 58 to 94 %. Regarding the total energy consumption of the HP it can be seen that for the parabolic reward function there are energy savings of around 13 % for all cases while for the step-shaped reward function there are energy savings of around 15 % for all cases compared to the hysteresis operation.

4.2. Operation of RL agent

The results shown in this section are based on the respective run with the lowest total electricity costs. Fig. 9 shows the agents actions on the HP as well as the behavior of the HWT for a week with a high heat demand. The same analysis for a week with a low heat demand can be seen in Fig. 10. 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

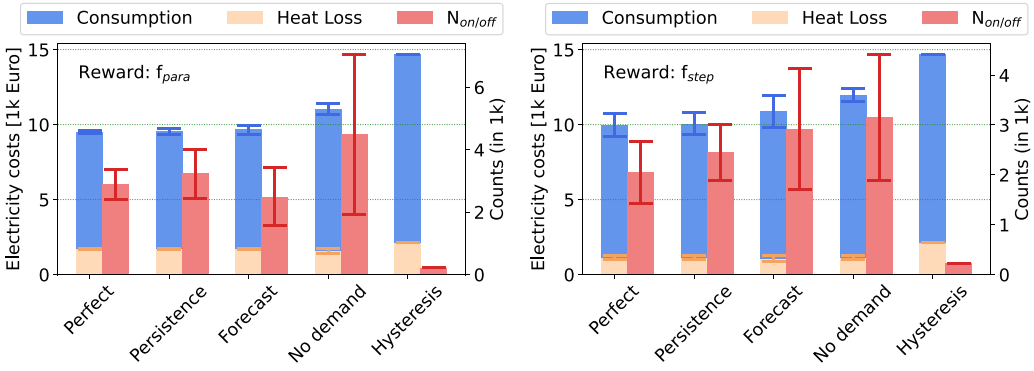


Fig. 7. Total electricity costs of the optimized agent using a parabola as the reward function (left) and a step function as the reward function (right) for the four different environments and the hysteresis operation. The total electricity costs are divided in costs due to consumption (blue) and costs due to heat loss (yellow).

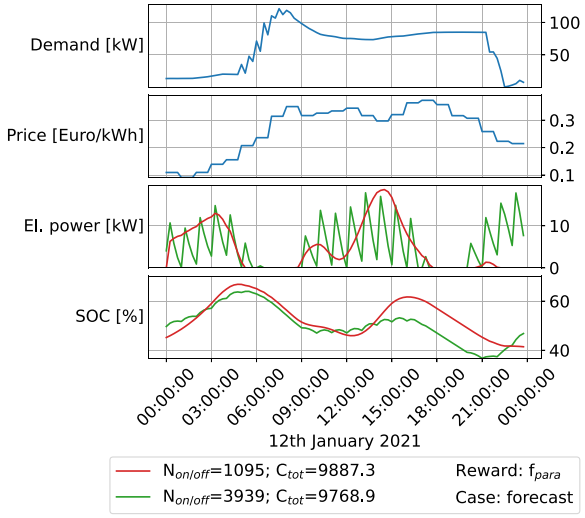


Fig. 8. Operation of two different policies for the case *forecast* with the parabolic reward function. (Time interval: 15 min)

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

	P_{avg} [kW]	P_{max} [kW]	SOC_{avg} [%]	SOC_{max} [%]	E_{tot} [kWh]
Hysteresis	100.0(0.0)	100.0(0.0)	55.0(0.0)	100.0(0.0)	20475(0)
			Reward function: parabola		
Perfect	9.5(1.2)	36.2(5.1)	46.1(0.9)	83.1(4.4)	17782(48)
Persistence	9.9(1.4)	41.1(9.4)	46.0(0.8)	87.3(6.5)	17775(69)
Forecast	8.3(1.4)	28.7(4.3)	45.8(1.0)	88.2(8.7)	17792(68)
No demand	7.7(2.5)	39.8(18.5)	41.5(4.5)	69.2(6.9)	17731(92)
			Reward function: step function		
Perfect	9.0(2.6)	30.2(8.4)	30.9(3.4)	89.5(13.8)	17492(96)
Persistence	9.6(1.5)	34.4(7.4)	30.4(3.4)	94.0(4.7)	17473(115)
Forecast	7.3(1.4)	24.2(3.9)	29.4(6.5)	80.0(16.8)	17492(118)
No demand	5.3(0.4)	22.8(5.4)	28.4(3.6)	58.1(10.7)	17436(96)

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 electricity costs compared to the hysteresis operation. It can also be observed in Fig. 10 that the electrical 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 Eq. (7). 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 Eq. (5) and Eq. (6) is not exceeded. A closer look at these fluctuations can be seen in Fig. 11.

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 electricity costs compared to a conventional hysteresis operation. The electricity cost 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 [41]. 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 [42]. Note, that we used scaled electricity prices from SMARD [27] which might overestimate the price fluctuations in the residential sector since additional grid fees and charges are not considered there. Thus, with more realistic residential electricity prices, the reported electricity cost savings might be lower. In future work, one could therefore rerun the simulations with different electricity price time series. However, the electricity cost savings are not only due to exploitation of price fluctuations but also due to pure energy savings. We observe energy savings of 13 % with the parabolic reward function and 15 % with the step-shaped reward function. Since the heat demand from the apartment complex is a fixed time series, these energy consumption savings arise from reducing losses within the HWT and shifting the operation of the HP to times with higher COPs. The RL agents are mainly able to reduce losses since they keep the SOC of the HWT lower compared to the hysteresis operation. The step-shaped reward function reaches higher loss savings due to the fact that the SOC is tried to be kept at an even lower level by the RL agent compared to the parabolic reward function.

Reviewing our approach we find that the parabolic reward function slightly outperforms the step-shaped reward function by means of total electricity 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 [43] 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.

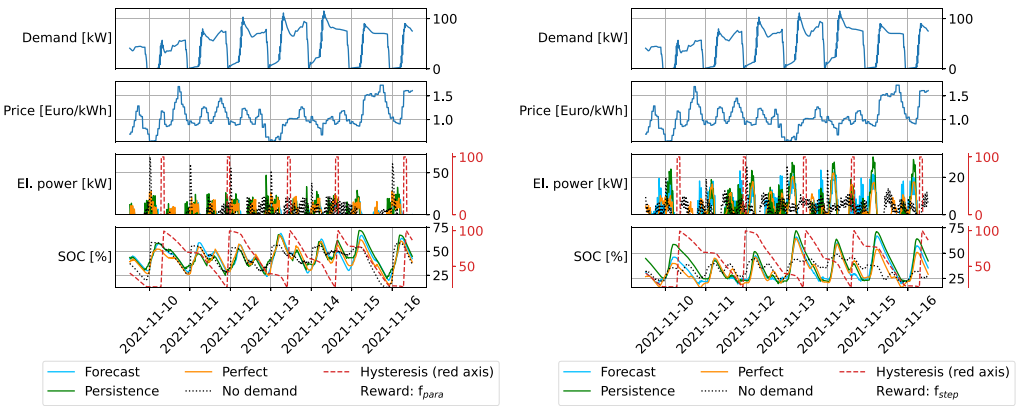


Fig. 9. Operation of the HP with the optimized agent on a week with a high heat demand using a parabola as the reward function (left) and a step function as the reward function (right) for the four different cases and the hysteresis operation. (Time interval: 15 min)

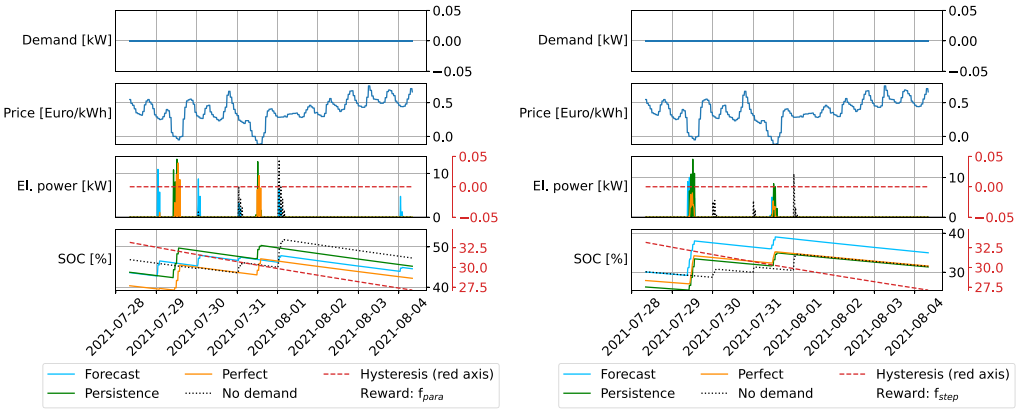


Fig. 10. Operation of the HP with the optimized agent on a week with no heat demand using a parabola as the reward function (left) and a step function as the reward function (right) for the four different cases and the hysteresis operation. (Time interval: 15 min)

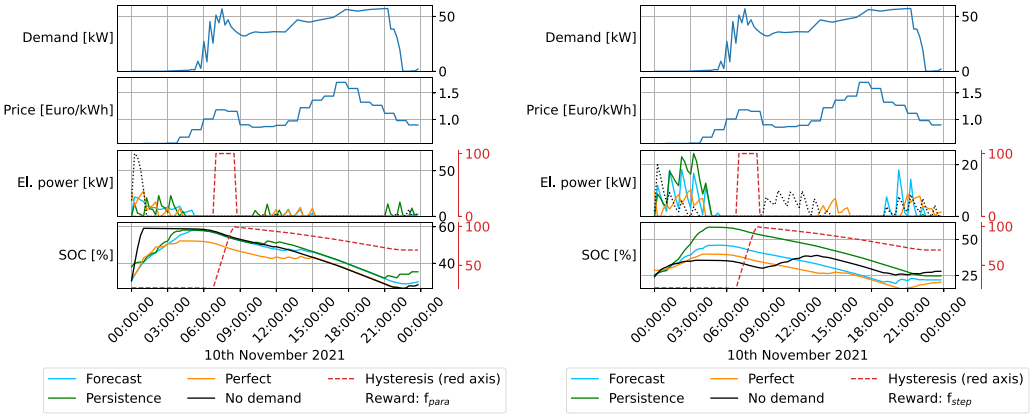


Fig. 11. Operation of the HP with the optimized agent on a winter day using a parabola as the reward function (left) and a step function as the reward function (right) for the four different cases and the hysteresis operation. (Time interval: 15 min)

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 [44] and [45] 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 electricity cost savings result to 24.7% and for the step-shaped reward functions the electricity cost savings are 18.5%. This is due to the installed HWT 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 electrical power used by the agent it gets clear that the installed nominal electrical power of the apartment complex's HP is not fully utilized. Dependent on the RL agent's setup, only 23–41% of the heat pump's installed nominal electrical 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 electrical power would only affect the hysteresis operation in the first place as long as the nominal electrical power is greater than the maximum electrical 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 and the apartment complex and the heat transfer between the HP and the HWT have been assumed to be loss free. While for the purpose of this work it is a valid assumption one should incorporate this loss in future work. If this cannot 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 at the heat exchangers between the apartment complex and the hot water tank to always be 5 °C in order to determine the SOC of the HWT.

This value of the temperature spread is a realistic assumption for a residential floor heating system, however the fixed spread is a simplification. In a real world system, small fluctuations of the spread are to be expected which is dependent on the heat demand. Theoretically, the spread can be fixed by setting the flow velocity which is called hydraulic levelization but this is rarely done. However, the fluctuations of a real world heating system could easily be included into the model by measuring the temperature of the respective flow when transferring the approach to a real building.

In our work, we model one central HWT with only one big water volume while in a real apartment complex, this central volume would likely be divided into multiple sub-volumes in close proximity to each other. They would share a common operational strategy and can be seen as a single heat storage. However, the heat losses via the surface of the hot water tank would be higher, since the surface of multiple smaller volumes is higher compared to one big volume. Additionally, the ambient temperatures of the sub-volumes might differ due to different positioning which also influences the losses. Currently, we assume a constant ambient temperature of the hot water tank of 20 °C. This could be elaborated in future work.

We also did not 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 HWT. Note, that we assume a simple heat pump model based on the regression of temperature-dependent COP values and the possibility to set the electrical operational power anywhere between 0 kW and 100 kW. We neglect the COP dependence on the electrical operational power of the HP. Given these assumption and procedure, our heat pump model is a simplification and idealization and could be elaborated in future work.

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 electricity cost savings and 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 h for one year of data where the HWT 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 apartment complex. It has been discovered that such an approach can significantly reduce electricity costs by approximately 35 % by exploiting the variations of a variable electricity price and by reducing the total energy consumption of the heat pump by up to 15 %. Additionally, we show that the intelligent operation of HPs does not use the full installed nominal electrical 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 HWT 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 HWT to minimize electricity costs especially due to losses in the HWT. 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, electricity costs are very similar for each run.

Improvements for further studies could be to increase the complexity of the environment. In Section 3 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 HWTs can be considered by expanding the observation and action space of the RL agent.

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 electrical power needed for the HP will roughly double.

CRediT authorship contribution statement

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

Data availability

The authors do not have permission to share data.

Acknowledgment 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). We would like to thank Patrik Schönfeldt for offering his expertise in the field of residential heating.

References

[1] Federal Environment Agency Germany (Umweltbundesamt). Energy consumption by energy source and sector. 2021, availabe at: https://www.umweltbundesamt.de/daten/energie/energieverbrauch-nach-energietraegern-sektoren#allgemeine-entwicklung-und-einflussfaktoren.
[2] Federal Environment Agency Germany (Umweltbundesamt). Energy consumption of private households. 2021, availabe at: https://www.umweltbundesamt.de/daten/private-haushalte-konsum/wohnen/energieverbrauch-privater-haushalte#hochster-anteil-am-energieverbrauch-zum-heizen.
[3] Deason J, Borgeson M. Electrification of buildings: Potential, challenges, and outlook. Curr Sustain/Renew Energy Rep 2019;6:131–9.
[4] Kazmi H, D'Oca S. 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. IEEE; 2016, p. 1–6.
[5] Mbiydzenyuy G, Nowaczyk S, Knutsson H, Vanhoudt D, Brage J, Calikus E. Opportunities for machine learning in district heating. Appl Sci 2021;11(13):6112.
[6] Noye S, Martinez RM, Carnieletto L, De Carli M, Aguirre AC. A review of advanced ground source heat pump control: Artificial intelligence for autonomous and adaptive control. Renew Sustain Energy Rev 2022;153:111685.
[7] Ntakolia C, Anagnostis A, Moustakidis S, Karcanias N. Machine learning applied on the district heating and cooling sector: A review. Energy Syst 2021;1–30.
[8] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy 2020;269:115036.
[9] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. J Artif Intell Res 1996;4:237–85.
[10] Mason K, Grijalva S. A review of reinforcement learning for autonomous building energy management. Comput Electr Eng 2019;78:300–12.
[11] Peirelinck T, Ruelens F, Decnoninck G. Using reinforcement learning for optimizing heat pump control in a building model in Modelica. In: 2018 IEEE international energy conference. IEEE; 2018, p. 1–6.
[12] Yuan X, Pan Y, Yang J, Wang W, Huang Z. Study on the application of reinforcement learning in the operation optimization of HVAC system. In: Building simulation, vol. 14, Springer; 2021, p. 75–87.
[13] Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy 2021;229:120725.
[14] Correa-Jullian C, Droguett EL, Cardemil JM. Operation scheduling in a solar thermal system: A reinforcement learning-based framework. Appl Energy 2020;268:114943.
[15] Lissa P, Schukat M, Keane M, Barrett E. Transfer learning applied to DRL-based heat pump control to leverage microgrid energy efficiency. Smart Energy 2021;3:100044.
[16] Ruelens F, Iacovella S, Claessens BJ, Belmans R. Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. Energies 2015;8(8):8300–18.
[17] Mbuwir BV, Geysen D, Spiessens F, Deconinck G. Reinforcement learning for control of flexibility providers in a residential microgrid. IET Smart Grid 2020;3(1):98–107.
[18] Heidari A, Marechal F, Khovalyg D. An adaptive control framework based on reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems. J Phys: Conf Ser 2021;2042(1):012006.
[19] Pujić D, Jelić M, Batić M, Tomašević N. Application of reinforcement learning for control of heat pump systems. 2022, p. 71–4.
[20] Langer L, Volling T. A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems. Appl Energy 2022;327:120020.
[21] Hummel S, Betzold C, Dentel A. Impact of the weather forecast quality on a mpcdriven heat pump heating system. In: CLIMA 2022 conference. 2022, http://dx.doi.org/10.34641/clima.2022.152.
[22] Dreher A, Bexten T, Sieker T, Lehna M, Schütt J, Scholz C, et al. AI agents envisioning the future: Forecast-based operation of renewable energy storage systems using hydrogen with deep reinforcement learning. Energy Convers Manage 2022;258:115401.
[23] Generische Gebäudesimulation als Bestandteil der Quartier-Simulationssoftware "QuaSi"-Verbundvorhaben EnStadtEs-West: Klimaneutrales Stadtquartier Neue Weststadt Esslingen. 2020.
[24] DIN eV. DIN 4108-6, Wärmeschutz und Energie-Einsparung in Gebäuden - Teil 6: Berechnung des Jahresheizwärme- und des Jahresheizenergiebedarfs. 2003.
[25] Crawley D, Pedersen C, Lawrie L, Winkelmann F. EnergyPlus: Energy simulation program. ASHRAE J 2000;42:49–56.
[26] Open Data Platform of Deutscher Wetterdienst, available at:, 2022. https://www.dwd.de/DE/leistungen/opendata/opendata.html.
[27] SMARD. Market Data, available at:, 2023. https://www.smard.de.
[28] Federal Network Agency Germany (Bundesnetzagentur). 2023. available at: https://www.bundesnetzagentur.de.
[29] ENTSO-E. European network of transmission system operators for electricity, available at:, 2023. https://www.entsoe.eu.
[30] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, arXiv:1707.06347.
[31] Watkins CJCH, Dayan P. Q-learning. Mach Learn 1992;8(3):279–92. http://dx.doi.org/10.1007/BF00992698.
[32] Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-Baselines3: Reliable reinforcement learning implementations. J Mach Learn Res 2021;22(268):1–8.
[33] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. OpenAI gym. 2016, arXiv:arXiv:1606.01540.
[34] J. Gerster KS, Sonnenschein M. Intelligentes Heimenergiemanagement – Nutzung der Synergiepotentiale bei der thermischen und elektrischen Objektversorgung durch modellbasierte und prädiktive Betriebsführungsstrategien, englisch: Intelligent home energy management - utilizing the synergy potential of thermal and electrical property supply through model-based and predictive operational management strategies. In: VDE-congress 2016 - internet of things. Mannheim, Germany; 2016.
[35] Kasturi P, Schwarz JS. Mosaik-heatpump. 2023, https://gitlab.com/mosaik/components/energy/mosaik-heatpump.
[36] Daikin. Commercial Heatpumps. 2024, available at: https://www.daikinapplied.eu/de/technologies/k%C3%A4ltemaschinen-und-w%C3%A4rmepumpen/w%C3%A4rmepumpen.
[37] Klement P, Schönfeldt P, Schmeling L. 3 multi-objective design optimisation of district energy supply – The influence of different domestic hot water concepts. In: Innovations and challenges of the energy transition in smart city districts. Berlin, Boston: De Gruyter; 2024, p. 35–52. http://dx.doi.org/10.1515/9783110777567-003.
[38] Schmeling L, Schönfeldt P, Klement P, Vorspel L, Hanke B, von Maydell K, et al. A generalised optimal design methodology for distributed energy systems. Renew Energy 2022;200:1223–39. http://dx.doi.org/10.1016/j.renene.2022.10.029.
[39] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. New York, NY, USA: Association for Computing Machinery; 2019, p. 2623–31. http://dx.doi.org/10.1145/3292500.3330701.
[40] Stable-Baselines3 Contributors. Stable Baselines3 documentation release 1.5.0. 2023, available at: https://stable-baselines3.readthedocs.io/_/downloads/en/v1.5.0/pdf/.
[41] Han G, Joo H-J, Lim H-W, An Y-S, Lee W-J, Lee K-H. Data-driven heat pump operation strategy using rainbow deep reinforcement learning for significant reduction of electricity cost. Energy 2023;270:126913. http://dx.doi.org/10.1016/j.energy.2023.126913.
[42] Peirelinck T, Ruelens F, Decnoninck G. Using reinforcement learning for optimizing heat pump control in a building model in Modelica. In: 2018 IEEE international energy conference. 2018, p. 1–6. http://dx.doi.org/10.1109/ENERGYCON.2018.8398832.
[43] Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of low exergy buildings. Appl Energy 2015;156:577–86. http://dx.doi.org/10.1016/j.apenergy.2015.07.050.
[44] Xue P, Jiang Y, Zhou Z, Chen X, Fang X, Liu J. Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms. Energy 2019;188:116085. http://dx.doi.org/10.1016/j.energy.2019.116085.
[45] Bünning F, Heer P, Smith RS, Lygeros J. Improved day ahead heating demand forecasting by online correction methods. Energy Build 2020;211:109821. http://dx.doi.org/10.1016/j.enbuild.2020.109821.