energies

MDPI

Article

Control of a PVT-Heat-Pump-System Based on Reinforcement Learning–Operating Cost Reduction through Flow Rate Variation

Daniel John * and Martin Kaltschmitt

Institute for Environmental Technology and Energy Economics, Hamburg University of Technology, Eissendorfer Strasse 40, 21073 Hamburg, Germany; kaltschmitt@tuhh.de
* Correspondence: daniel.john@tuhh.de; Tel.: +49-40-42878-3322

Abstract: This study aims to develop a controller to operate an energy system-consisting of a photovoltaic thermal (PVT) system combined with a heat pump, using the reinforcement learning approach to minimize the operating costs of the system. For this, the flow rate of the cooling fluid pumped through the PVT system is controlled. This flow rate determines the temperature increase of the cooling fluid while reducing the temperature of the PVT system. The heated-up cooling fluid is used to improve the heat pump's coefficient of performance (COP). For optimizing the operation costs of such a system, first an extensive simulation model has been developed. Based on this technical model, a controller has been developed using the reinforcement learning approach to allow for a cost-efficient control of the flow rate. The results show that a successfully trained control unit based on the reinforcement learning approach can reduce the operating costs with an independent validation dataset. For the case study presented here, based on the implemented methodological approach, including hyperparameter optimization, the operating costs of the investigated energy system can be reduced by more than 4% in the training dataset and by close to 3% in the validation dataset.

Keywords: PVT; reinforcement learning; solar-assisted heat pump; control approaches; operating cost analysis

check for updates

Citation: John, D.; Kaltschmitt, M. Control of a PVT-Heat-Pump-System Based on Reinforcement Learning–Operating Cost Reduction through Flow Rate Variation. *Energies* **2022**, *15*, 2607. https://doi.org/10.3390/en15072607

Academic Editor: Lyes Bennamoun

Received: 2 March 2022
Accepted: 31 March 2022
Published: 2 April 2022

1. Introduction

Photovoltaic-thermal (PVT) systems-an extension of "classical" photovoltaic (PV) systems-can provide both electrical energy and low-temperature heat at different temperature levels, depending on the given environmental parameters, such like current solar irradiation and ambient temperature. By varying the flow rate of the cooling fluid pumped through the cooling structure of the PVT system, the temperature level of this cooling fluid can be controlled. Consequently, the efficiency of an attached heat pump assumed to be installed together with such a PVT systems can be increased. As a consequence, the provided thermal energy of a system or building can potentially be provided at lower operating energy costs. To be able to better utilize and incorporate the advantages of PVT systems into energy systems is of great relevance for the impact of PVT systems.

Since both the environmental parameters (e.g., solar radiation, ambient temperature) as well as the demand characteristics (e.g., space heating, domestic hot water, electricity) for energy systems are typically volatile, "classical" on-off control concepts of the flow rate of PVT systems cannot explicitly respond to fluctuations in energy supply and demand in a most cost-efficient way. In parallel, variable flow rates can add value to PVT systems [1–3]. Thus, a controller development is necessary to realize variable flow rates in a most cost-efficient way. Therefore, a reinforcement learning (RL) approach is applied to control the flow rate related to the volatile energy supply and demand characteristics.

This approach aims to enable a control decision regarding an advantageous flow rate depending on measurable variables of the combined PVT-heat pump-system under investigation.

In such a reinforcement learning approach, measurable variables or observations of a system are collected and transmitted to a control unit called "agent". This agent uses,

in most cases, a trained artificial neural network (ANN) to select an appropriate action based on these observations and returns this selection to the system. After performing the respective action, the agent receives a feedback, called a reward, from the system. The ANN corresponds to the agent's policy being the core of the RL approach. The goal of this RL approach is to enable the agent to choose a beneficial action for the particular system under consideration (i.e., to maximize the reward). This goal can be achieved by an extensive training of the ANN and thus the agent's policy. The advantage of such an approach is that no prior knowledge about the system to be controlled is needed and by trial and error a policy for maximizing the reward can be learned.

Existing Studies

The technology of PVT systems has already existed for several decades [4,5]. However, it was not until the sharp price drop in the module prices for photovoltaic (PV) systems in the past decade that the installation and application of PVT systems moved into a realistic economic range to gain step by step increasingly higher market shares.

Several review papers summarized the given possibilities to vary the performance of PVT systems [6–11]. Possible improvements in electrical and/or thermal yield through adjustments in cooling structures [12–16] or by varying the cooling fluid [17,18] have also been investigated. Some studies also addressed the impact of the flow rate on the performance of PVT systems [1], through absorber type variation [19,20], the addition of phase change materials [21], the usage of nanofluids [2] or the combination with domestic hot water tanks [22] or hydrogen production [23]. Such a variable flow rate in PVT systems shows two possible effects: first, an increased electrical yield of the PVT system and second, a controllable temperature level of the cooling fluid [24].

Individually, both effects are typically of minor importance for the overall performance of a PVT system. However, in combination with a heat pump assumed here, the temperature level of the cooling fluid might significantly influence the overall system performance [25]. This is due to the fact that the heat pump's coefficient of performance (COP) increases by raising the temperature level on the cold side using low-temperature heat from the PVT system (i.e., the heated up cooling fluid) [26,27].

Some studies have also considered combined PVT-heat-pump-systems. A fixed flow rate with a "classical" on-off control has mainly been realized within these systems. The results of these papers [28–31] consistently show positive results for the interaction of the two technologies. It is suspected that variable flow rate control can further increase the effectiveness of the interaction between PVT and heat pumps.

Reinforcement learning (RL) is a possible control approach. Ever since the success of AlphaGo in 2015 [32], due to the increased computational capacity available in recent years, this machine learning approach has generated increased interest. Even though this method is most commonly used for gaming and image processing, initial research on the application of RL approaches in energy systems has been published in recent years.

One study linking the reinforcement learning approach to PVT systems has taken control of energy flows between the individual components of an energy system showing positive results [33]. Other studies have also assessed the potential of the RL approach for control tasks in various energy systems (e.g., [34–38]). However, some studies show the risk of affecting comfort for occupants of the respective buildings studied. A flow rate control of PVT systems based on the RL approach has not yet been investigated.

To sum up, the current literature shows promising approaches for PVT systems with a variable flow rate and their potentials in integrated energy systems. This is especially true for the joint installation of a PVT system with a heat pump. Approaches for an automated control of the flow rate of PVT systems indicate a certain potential as well. An adaptive method such as reinforcement learning can be the link for an efficient integration of an automated control of such a PVT system in various applications, especially under individual as well as stochastic conditions. Thus, the combination of these two topics—PVT and reinforcement learning—promises to be an exciting area of research. Consequently, this

paper investigates a reinforcement learning approach for flow rate control in an integrated PVT-heat-pump system.

2. Methodological Approach

This paper aims to present a methodological approach to train a controller, based on the reinforcement learning (RL) approach, to reduce the operating costs of a PVT-heat-pump system by varying the flow rate through the PVT collectors and comparing to a reference controller. Accordingly, the objective can be divided into two parts; one part is the development of a controller based on reinforcement learning and the other part is the demonstration and quantification of the operational cost reduction using the previously developed controller. The methodological approach to achieve this objective is presented in Figure 1.

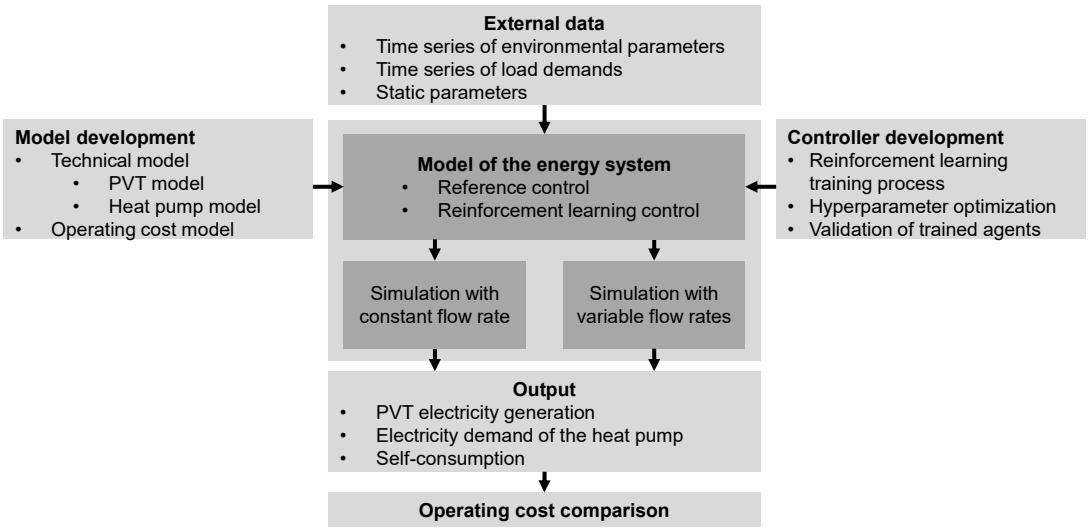

Figure 1. Methodological approach to compare operating costs between a reference control and a reinforcement learning control.

The methodological approach is divided into the following steps.

- First, the energy system under investigation (here: a combination between a PVT system and a heat pump) is modelled. Then, a reference control with a constant flow rate is implemented into this model and the overall energy system is simulated. Afterward, the operating costs for this system to meet the electricity, space heating and domestic hot water demands are determined.
- Second, to develop a controller for a system operation with variable flow rates, the reinforcement learning agents are trained and the respective hyperparameters are determined using Bayesian optimization. The trained reinforcement learning agents are afterwards validated with a new dataset of external data.
- Third, a trained agent is implemented as a control unit and the overall model is simulated again. Then, the operating costs for the simulation results based on the reinforcement learning based control to meet the electricity, space heating and domestic hot water demands are determined.
- Finally, the operating costs for the simulation are compared with a reference control and with the reinforcement learning based control. If the reinforcement learning agent ensures lower operating costs than are achieved with the reference control, the entire methodological approach is evaluated as successful.

Below, the model development of the PVT-heat-pump-system is presented first. Then the controller development with the reinforcement learning approach and the hyperparameter optimization is described.

2.1. Model Development

The energy system under investigation is divided into a technical model consisting of a PVT model and a heat pump model as well as an operating cost model. The overall system model was designed as a quasi-stationary model with variable time step sizes using commercial software [39]. The overall system model is shown in Figure 2.

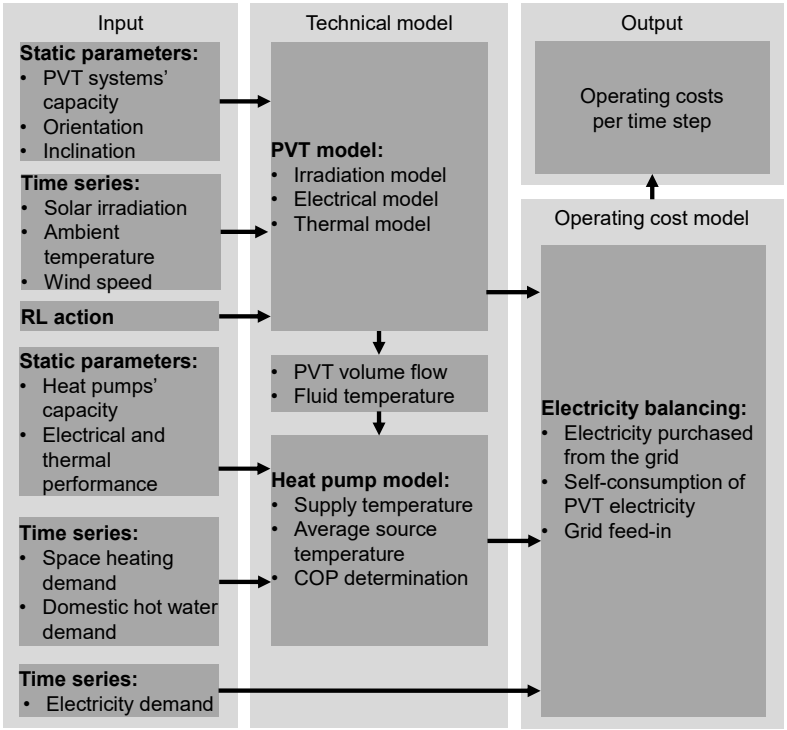

Figure 2. Overall energy system model consisting of the technical model with a PVT model and a heat pump model as well as the operating cost model.

2.1.1. Technical Model

The technical model represents a PVT and a heat pump model including their interaction. The interaction consists of the increase of the temperature level on the cold side of the heat pump, due to the temperature increase of the cooling fluid from the PVT system, as well as the volume flow available at this temperature level.

PVT Model

The PVT model simulates the operational behavior of a PVT system (for details see [24]). It consists of three sub-models, an irradiation model, an electrical model and a thermal model. This model approach was validated based on measured data from a test stand for PVT collectors. The required input parameter as time series are solar irradiance, ambient temperature, and wind speed. Static parameters such as the capacity of the PVT system and the orientation and inclination are also entered into the model. The outputs are the PVT collectors' electrical power, the volume flow of all PVT collectors, and the respective fluid temperature. By modeling the thermal inertia of the PVT collector, different time step sizes can be simulated.

Heat Pump Model

The model describes a commercially available brine-to-water heat pump in the single-digit kW-range for space heating and domestic hot water supply for single-family houses [40]. The input values are the volume flow of all PVT collectors of the system, the fluid temperature after passing through the PVT collectors, the given brine temperature and the respective

space heating and/or domestic hot water demand. Electrical power is required to operate this heat pump to meet these energy demands.

For space heating demand, the supply temperature demanded by the heating circuit to meet the given heat demand depends on the ambient temperature and is calculated according to Equation (1). $T_{heat,supply}$ equals the supply temperature of the heating circuit and $T_{ambient}$ describes the ambient temperature. The gradient of this dependency of the supply temperature on the ambient temperature is assumed to be 0.4.

$$T_{heat,supply} = 35\ ^{\circ}\mathrm{C} - 0.4^{\circ} T_{ambient} \quad (1)$$

The heat pump's electrical and thermal performance values for different source and sink temperatures, provided by the data sheet of the simulated heat pump, are included as parameters in the model. This is followed by a calculation of the average source temperature for the required mass flow of the heat pump ($T_{heat\ pump,source}$) by mixing the PVT flow rate (m_{PVT}) with the corresponding fluid temperature ($T_{PVT,out}$) and the assumed brine temperature (T_{brine}) with the additional required mass flow of brine (m_{brine}), provided via borehole heat exchangers (Equation (2)).

$$T_{heat\ pump,source} = \frac{T_{PVT,out}{}^{\circ}\dot{m}_{PVT} + T_{brine}{}^{\circ}\dot{m}_{brine}}{m_{PVT} + m_{brine}} \quad (2)$$

Since the required volumetric flow rate on the cold side of the heat pump depends on the average source temperature resulting from mixing, an iterative procedure is used to determine both the average source temperature and the required volumetric flow rate of the heat pump. At the same time, the electrical and thermal power of the heat pump at the respective time step is determined according to the average source temperature resulting from mixing and the corresponding coefficient of performance (COP).

An on-demand approach is used to cover the demand for space heating and domestic hot water in the energy system model considered here. Therefore, the heat pump must provide the demanded thermal outputs in each time step.

2.1.2. Operating Costs Model

In the operating cost model, the ongoing operating costs per time step are calculated from the point of view of the operator of the PVT-heat-pump-system resp. the building where this system is integrated, i.e., the own use of the PVT electricity is evaluated positively, since this reduces the amount of electricity purchased from the public grid. In this model, only the operating costs are considered, i.e., the electricity production costs of the PVT electricity are assumed to be zero. The grid electricity tariff, the heat pump tariff (if available) and the feed-in tariff (if available) for the PVT electricity are needed as external input parameters.

In the first step of the operating cost model, electricity production and demand are balanced. Electricity provided by the PVT system is used according to the following priority.

- Circulation pump to set the flow rate falls under the PVT system's own demand, as meeting this demand is necessary to gain the low-temperature heat.
- Electricity demand of the system/building, as meeting this demand is usually more expensive than meeting the electricity demand of the heat pump.
- Electricity demand of the heat pump.
- Electricity sold to the public grid.

The respective electricity demands are multiplied proportionally by the assumed tariffs to obtain the ongoing operating costs (Equation (3)).

$$C_{operating} = C_{Grid}\ E_{Grid,Elec.} + (C_{Feed-in} - C_{Grid})\ E_{PVT,Elec.} + C_{HP}\ E_{Grid,HP} + (C_{Feed-in} - C_{Grid})\ E_{PVT,HP} - C_{Feed-in}\ E_{PVT,surplus} \quad (3)$$

C_{Grid} equals the electricity price for purchase from the public grid, $C_{Feed-in}$ describes the feed-in tariff, and C_{HP} represents the heat pump electricity tariff. In addition, $E_{Grid,Elec.}$ equals the electricity purchase from the grid to cover the electricity demand, $E_{PVT,Elec.}$ describes the self-consumption of the PVT electricity, $E_{Grid,HP}$ represents the electricity purchase from the grid to cover the heat pump electricity demand, $E_{PVT,HP}$ equals the self-consumption of the heat pump PVT electricity, and $E_{PVT,surplus}$ describes the excess PVT electricity fed into the public grid.

The output are the operating costs of the respective time step. These serve as rewards for the reinforcement learning approach explained below. If the operating costs per time step are added up over all time steps, the operating costs of the respective episode under consideration are obtained, representing the benchmark for the reinforcement learning agents to be trained.

2.2. Controller Development

The controller development is divided into the reinforcement learning approach and its training process as well as the hyperparameter optimization using Bayesian optimization.

2.2.1. Reinforcement Learning

Reinforcement learning (RL) is a machine learning approach where an agent independently learns a policy through trial and error by maximizing rewards received. This approach is used to control the flow rate through the PVT collectors to reduce the operating costs of the overall system. Therefore, measurable variables (observations) of the modelled system are transferred to a control unit (agent). In this control unit, these measured variables are processed, and a decision (action) is made regarding the flow rate to be set for the PVT system. This decision on the flow rate influences the resulting operating costs and the state of the PVT-heat-pump-system. Thus, a value can be assigned to these changes representing the reward for the PVT-heat-pump system for the chosen decision depending on the previously measured variables of the system.

In the following, the used reinforcement learning (RL) terms are defined. Then the RL agent type is introduced, followed by a description of the components of its artificial neural network (ANN).

Figure 3 shows the main components of the reinforcement learning approach. The relevant components are the environment and the agent linked via the state, the action and the reward.

- Environment. The environment corresponds to the model of the overall energy system described earlier. Variables of this model are shared as states with the agent.
- State. The state describes the observed properties of the environment, passed to the agent as scalars. They form the input for the agent's ANN. For example, a possible state could be the ambient temperature.
- Agent. The agent represents the core of the reinforcement learning approach. It depicts the policy of the control system. By using an ANN as a function approximator of the control system, the states of the environment are mathematically transformed into an action as output of the ANN.
- Action. The action corresponds to the output value of the ANN or the agent. In the system under investigation, it is the selected flow rate for the PVT collectors.
- Reward. The reward describes the feedback of the environment to the agent after a selected action has been performed. An example is the negative operating cost per time step. The goal for the agent is to maximize the reward over a period of time steps called an episode (e.g., a month or a year).

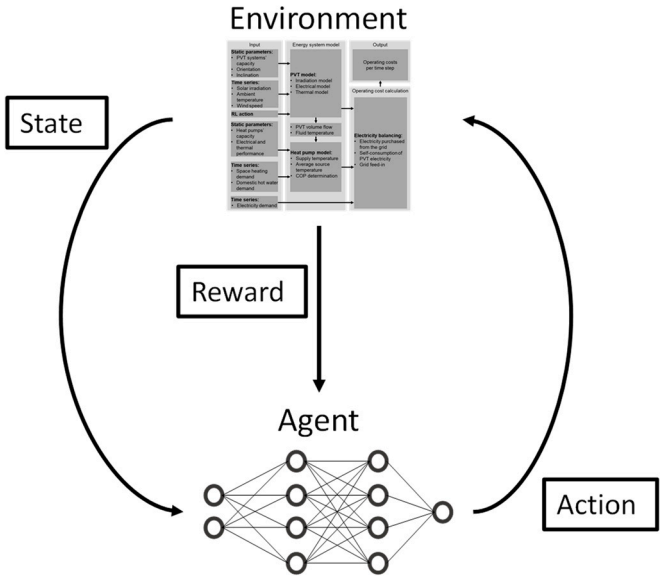

Figure 3. Main components and links of the reinforcement learning approach.

A variety of different agent types are available and are divided into policy-based and value-based agents. The agent type used here is the value-based Deep Q-Network (DQN) agent, which trains an ANN to estimate the reward of a selected action. The DQN agent, as a variant of Q-learning, was chosen due to its successful implementation in problems with complex and high-dimensional environments. Moreover, DQN agents are highly sample-efficient. A continuous observation space and a discrete action space are implemented for the DQN agent. This type of agent is characterized by the properties model-free, online and off-policy.

- Model-free means that the agent's reward can be explicitly computed.
- Online means that the agent learns from its own accumulated experience by interacting with the environment.
- Off-policy means that the agent can learn the optimal policy independently of the action currently being performed.

After each simulated time step, the tuple of the state, action, reward, and subsequent state is stored in the experience buffer. A random collection of such stored tuples, called mini-batches, is then taken from the experience buffer and used to update the weights of the ANN. Here, a one-step minimization of the loss function is performed over all the tuples of the mini-batch (Equation (4)). L is the loss function, which is the summed squared deviation over all M mini-batches from the value function target y, obtained as a reward from the environment, to the value function target Q, obtained by the existing ANN when applying state S and action A [41]. After updating the policy/weights of the ANN, the next time step is simulated until the end of a training episode is reached.

$$L = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(S_i, A_i|\phi))^2 \quad (4)$$

The ANN of the DQN agent in this study is composed of the Long Short-Term Memory (LSTM) layer and the Fully Connected layer.

- LSTM layers represent a special type of layer enabling the efficient processing of data from time series [42]. This type of layer is able to handle the data in the form of time series of environmental parameters as well as the demand curves of the investigated energy system.
- The Fully Connected layer represents the classical framework of an ANN. Different activation functions can basically be used in each layer of the ANN. Here, the Rectifier

Linear Unit (ReLU) is used, allowing negative values to pass on with zero and positive values to pass on unchanged.

2.2.2. Hyperparameter Optimization

Hyperparameters represent superordinate variables or settings remaining constant throughout an entire training process. Although these hyperparameters have an enormous influence on the training success of the respective agents, their values are often selected based on the user's experience. In this case, however, hyperparameter optimization is performed using Bayesian optimization [43,44], which are chosen because it is very efficient in evaluating a cost-intensive objective function (due to the training time of an agent of the RL approach). The objective function is hereby the cumulated reward over an episode (e.g., the operational costs).

Since the optimal hyperparameter set for the maximization of the objective function is not known, it must be approximated over the solution space of the hyperparameters. For this purpose, a model of the solution space (all variable hyperparameters) is created via a surrogate function. Based on the iterative determination of the objective function on a hyperparameter set, the expected maximum of the surrogate model is determined via an acquisition function suggesting promising hyperparameter configurations for evaluation. This point in the solution space is then taken for the next iteration (i.e., the next training), where each iteration of the Bayesian optimization trains one agent at a time. The goal here is to determine the values of each hyperparameter with which the training of the ANN is successful. The methodological procedure for the Bayesian optimization with the underlying levels of reinforcement learning training process and energy system simulation is shown in Figure 4.

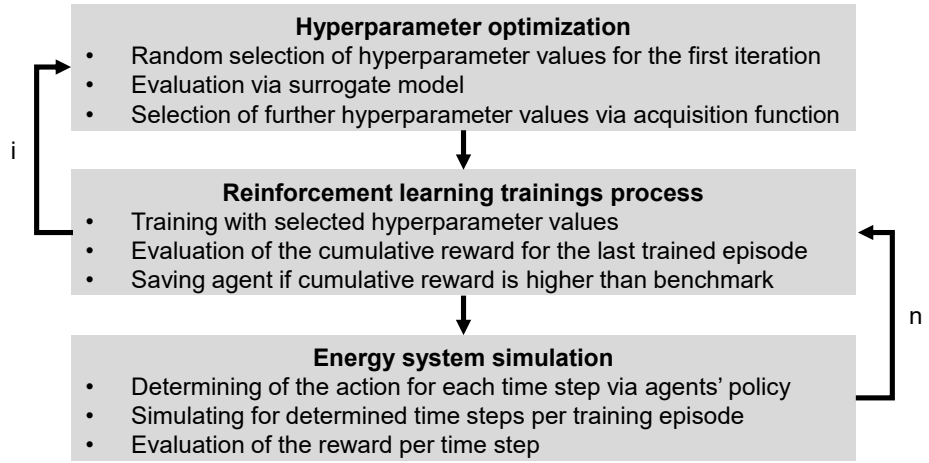

Figure 4. Methodological procedure for hyperparameter optimization, where i is the number of iterations for hyperparameter optimization and n is the number of training episodes of each reinforcement learning training process.

For the Bayesian optimization, all relevant adjustable hyperparameters are used. In this respect, it hyperparameters of the training process and hyperparameters of the utilized ANN (number of hidden units in applied layers and the section depths of fully con-nected layers) can be distinguished. All varied hyperparameters are briefly explained below.

- Number of training episodes describes over how many completed time periods (one episode) the ANN should be trained. This equals the stop criteria for each iteration.
- Number of hidden units in the LSTM layer equals the number of LSTM cells in the implemented LSTM layer here.
- Number of hidden units in each of the possible fully connected layers equals the number of neurons in each fully connected layer of the implemented ANN.
- Learning rate describes the respective factor for updating the weights of the neurons.

- Section depth specifies the number of fully connected layers for the ANN in each iteration step of the Bayesian optimization.
- Sequence length for the time series in the LSTM layer describes how many elements of the time series under consideration are used in the ANN.
- Size of the mini-batch describes how many tuples from state, action, reward and subsequent state are used for updating the loss function.
- Length of the experience buffer describes how many of these tuples are stored.
- Discount factor describes how much future rewards should be included in the current reward.

The Bayesian optimization is performed using a training dataset. The agents trained during the hyperparameter optimization are saved if their summed reward (which corresponds to the negative operating cost of the training episode) is higher than the reference operating cost of the training dataset.

Afterwards, the agents trained in this way are applied to a validation dataset without further training. Here, the annual operating costs are also determined and compared to the reference operating costs of the validation dataset. The training of the agents and the controller development based on reinforcement learning can be considered successful if trained agents show a reduction in the operating costs for both datasets used.

3. Framework and Datasets

First, the framework assumptions for the energy system under investigation are presented, then the reference control and reinforcement learning (RL) control framework, and finally the framework for controller development, including datasets for training and validating the RL control.

3.1. Definition of Energy System

Figure 5 shows the energy system under investigation with all relevant technologies. The circulation pump is the element to be operated by the reinforcement learning based control.

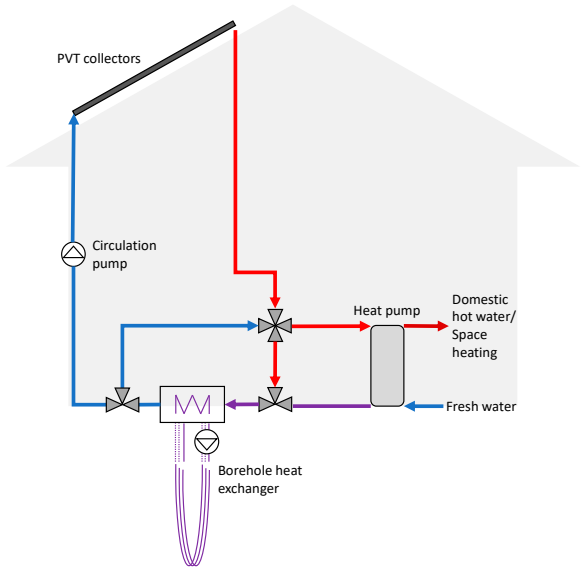

Figure 5. Energy system under investigation—Installation of PVT collectors, a brine-to-water heat pump and a borehole heat exchanger for the provision of a constant brine temperature. The circulating pump is actively operated by the two control approaches studied here.

The assumptions for the energy system shown in Figure 5 for the single-family house (SFH) assessed here are presented in Table 1. The selected values are typical for the size of PVT systems in SFH and correspond here to 20 PVT modules with 280 W electrical power

each. The single-family house assumed here represents an average SFH for Germany with a construction year of 2016 and after, with energy demand and the respective electricity tariffs selected accordingly [45]. The capacity of the heat pump is also selected to correspond to the energy demand of the SFH. All energy system assumptions made apply to both control approaches implemented in this work.

Table 1. Assumptions for the modeled energy system, for the single-family house (SFH), and electricity tariffs.

Energy System	Value	Unit	Single-Family House	Value	Unit	Tariffs	Value	Unit
PVT capacity installed	5.6	kW	Living space of the building	187	m^2	Public grid tariff	0.31	€
Orientation	0	°	Space heating per area	45	$\frac{kWh}{m^2\ a}$	Heat pump tariff	0.22	€
Inclination	30	°	Occupancy	3	Person	Feed-in tariff	0.07	€
Heat pump capacity installed	6	kW	Electricity demand per person	1750	$\frac{kWh}{Person\ a}$			
Brine temperature	10	°C	Domestic hot water demand per person	500	$\frac{kWh}{Person\ a}$			

3.2. Definition of Reference Control

For the reference control, static conditions are assumed to activate the circulation pump for the PVT collectors and maintain the cooling fluid's flow rate. That means that the circulation pump for the cooling fluid is switched on, when the solar radiation on the surface of the PVT module and the difference between the temperature of the PVT collectors and the inlet temperature of the cooling liquid exceed 3 K. If one of these two static conditions is not met, the circulation pump is switched off. If both static conditions occur, a constant flow of 30 L/h per PVT collector is set. This is a typical value for the flow rate of PVT collectors and typically shows a good balance between the electrical and thermal yield [28]. A constant brine temperature of 10 °C (ground water temperature) is assumed for the cooling fluid pumped through the PVT collectors and for mixing of the volume flow of the heat pump.

3.3. Definition of Reinforcement Learning (RL) Control

The control approach based on reinforcement learning does not impose any static conditions. The agent should select beneficial actions purely from the rewards provided by the modeled energy system. The action space consists of seven discrete flow rates between 0 L/h and 90 L/h in 15 L/h increments and it is assumed that a corresponding circulator pump can easily and immediately set the appropriate flow rates.

The agent's observations of the environment/model are shown in Table 2. Thus, environmental parameters such as solar irradiance and ambient temperature, electricity demand, and calendrical values are passed to the agent. The agent's artificial neural network processes these observations into an action and then receives the reward for this action from the environment. In this case, the reward function is given in Equation (3) of the operating cost model. The cumulative reward over all time steps of an episode then represents the decisive value for the evaluation of the reinforcement learning based control approach.

Table 2. Agent's observation of the environment.

Observation	Unit
Direct solar irradiation	$\frac{W}{m^2}$
Diffuse solar irradiation	$\frac{W}{m^2}$
Ambient temperature	°C
Wind velocity	$\frac{m}{s}$
Electricity demand	$\frac{kWh}{h}$
Hour of the day	–
Day of the year	–

3.4. Framework Controller Development

Based on the methodological approach, two different datasets are required for the development of the reinforcement learning control. These datasets are consequently called the training dataset and validation dataset. Both datasets consist of the same parameters as the agent observations (Table 2) and the hourly demand for space heating and domestic hot water.

The training dataset includes environmental parameters such as weather data for Hamburg, Germany from 2018 and synthetic load curves of the described single-family house and its electricity, space heating and domestic hot water demand. These load curves were generated using VDI 4655 representing average and thus smoothed load curves [46]. To increase the training speed, the data were taken according to the central week of each month of 2018 and the training was performed with this shortened, 12-week, period.

The validation dataset contains the same parameters as the training dataset. The weather data for the location Hamburg for the validation data set are from the year 2017. The environmental parameters for both years (2017 and 2018) are based on [47,48]. To obtain a difference of the load curves of the training dataset, the data for space heating and hot water demand for the validation dataset and the year 2017 were taken from [49]. The electricity demand comes from a dataset consisting of the actual measured electricity loads of 74 single-family homes also for the year 2017 [50]. These 74 electricity demand curves are averaged to create one validation case enabling the testing of the trained agents or control units on an independent load curve.

Bayesian optimization takes place on the training dataset and requires lower and upper bounds for the hyperparameters under study. These lower and upper boundaries of the varied hyperparameters are shown in Table 3. In the first iteration of Bayesian optimization, a hyperparameter set is randomly selected to begin evaluating the objective function.

Table 3. Hyperparameters and their respective lower and upper boundaries.

Hyperparameter	Lower Boundary	Upper Boundary
Learning rate	10^{-6}	10^{-2}
Number of training episodes	50	500
Number of hidden units in the LSTM layer	1	50
Number of hidden units in the fully connected layers	1	100
Section depth of fully connected layers	0	3
Sequence lengths	2	24
Mini-batch size	1	48
Experience buffer length	10^4	10^6
Discount factor	0	1

4. Results and Discussion

The results are presented divided into controller development results, i.e., Bayesian optimization results and operating cost results of trained agents, and results regarding the comparison of static reference control and reinforcement learning (RL) control in terms of operating costs.

4.1. Controller Development

A sample of 244 iterations of the Bayesian optimization was taken. The most important results are shown in Figure 6a, where the influence of the Bayesian optimization on the training success of the reinforcement learning approach can be seen. While the first iterations of the Bayesian optimization were performed with randomly selected hyperparameters, the success of the training increased as the evaluation of the objective function progressed. The objective function contains the operating cost over the training period equaling one episode of training. As a cost reference, an objective value of 132.3 was determined, shown as a horizontal line in Figure 6a.

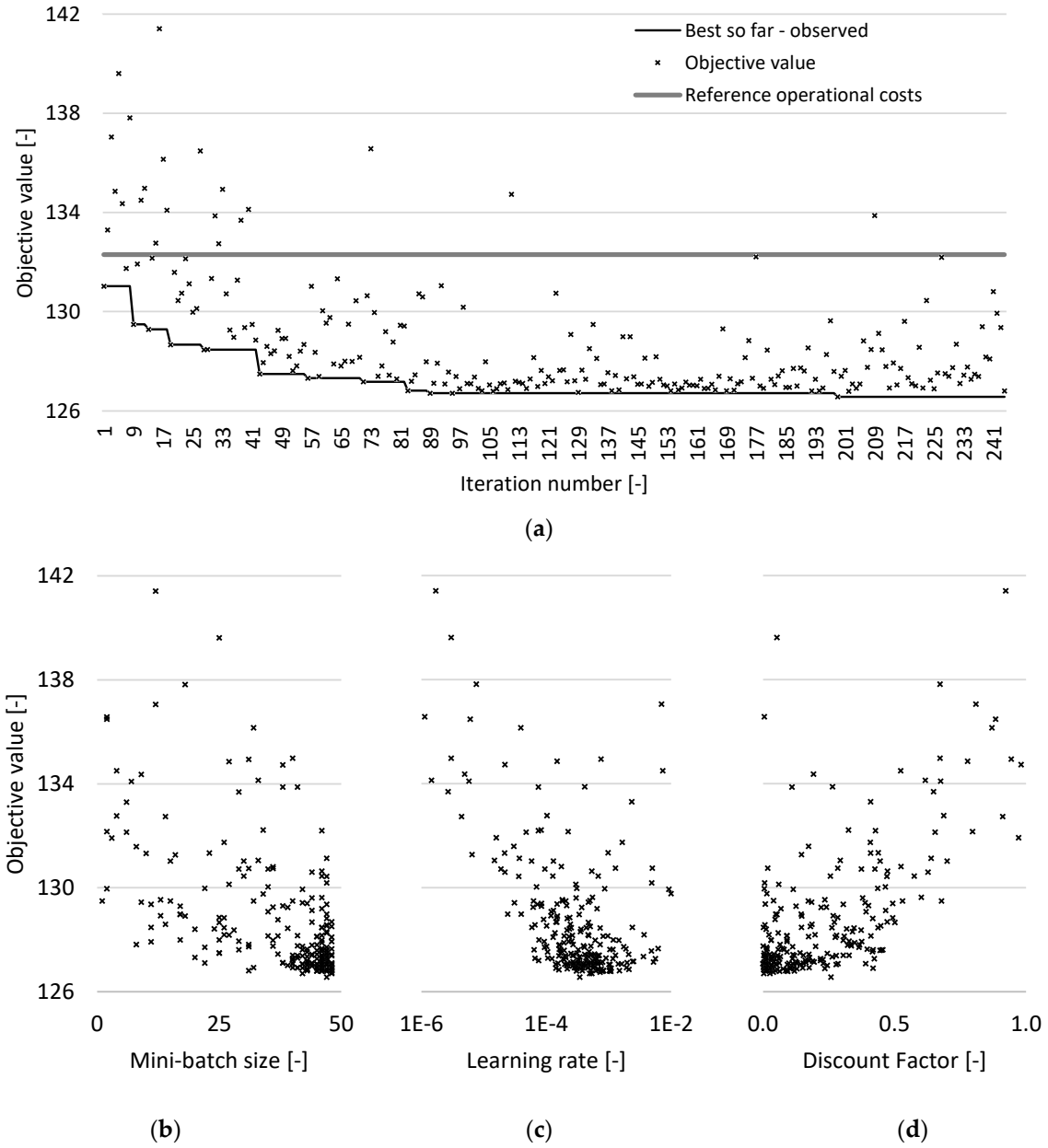

Figure 6. Results of the Bayesian optimization—Objective values over all iteration steps (**a**), effect of mini-batch size on objective value (**b**), effect of learning rate on objective value (**c**), and effect of discount factor on objective value (**d**).

The three hyperparameters, learning rate, size of the mini-batch and discount factor, are also shown in Figure 6b–d. These three hyperparameters clearly influence the minimization of the objective function (i.e., the operating costs), while the other investigated hyperparameters have no evident influence.

All trained agents were now applied to both datasets. Figure 7 shows a boxplot of the cost ratios of the operating costs of all trained agents applied to the training and validation datasets. Here, the operating costs of the trained agents were each divided by the previously defined reference operating costs to be able to define the cost ratios and subsequently the savings. For most parts, the trained agents show operating cost reductions. The cost ratios in the validation dataset are a little closer to the reference control. Nevertheless, operating cost reduction could be shown even when applied to the dataset unknown to the agents in the validation.

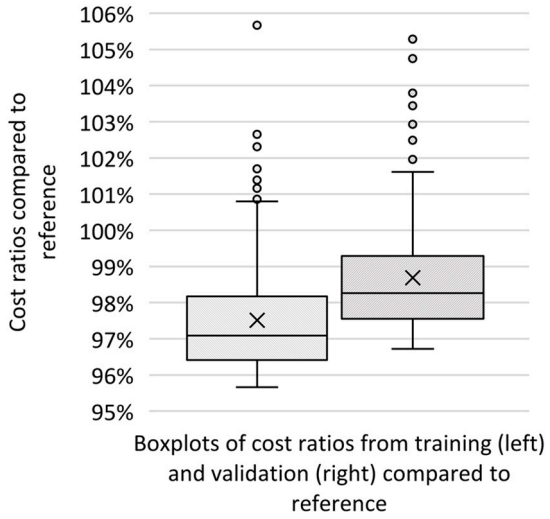

Figure 7. Boxplot of cost ratios of the operating costs applied on training and validation dataset.

Next, to realize the control approach based on reinforcement learning, the selection of one agent per system/building is necessary. In order to determine how likely an operating cost reduction of an agent is if this agent shows an operating cost reduction on the training dataset, the conditional probability according to Bayes' theorem is required.

Figure 8 shows the probabilities relevant for this, split according to the savings of the trained agents compared to the reference. The dotted lines show the shares of agents with a certain percentage of operating cost reductions. Here, $P(Train)$ represents the shares of trained agents per operating cost reduction, and $P(Train \cap Val)$ represents the shares of simultaneous operating cost reductions in the training and validation datasets. The bars now show the conditional probability $P(Val|Train)$; i.e., how likely it is that a trained agent will also show an operating cost reduction on the validation dataset if it has already shown certain operating cost reductions on the training dataset.

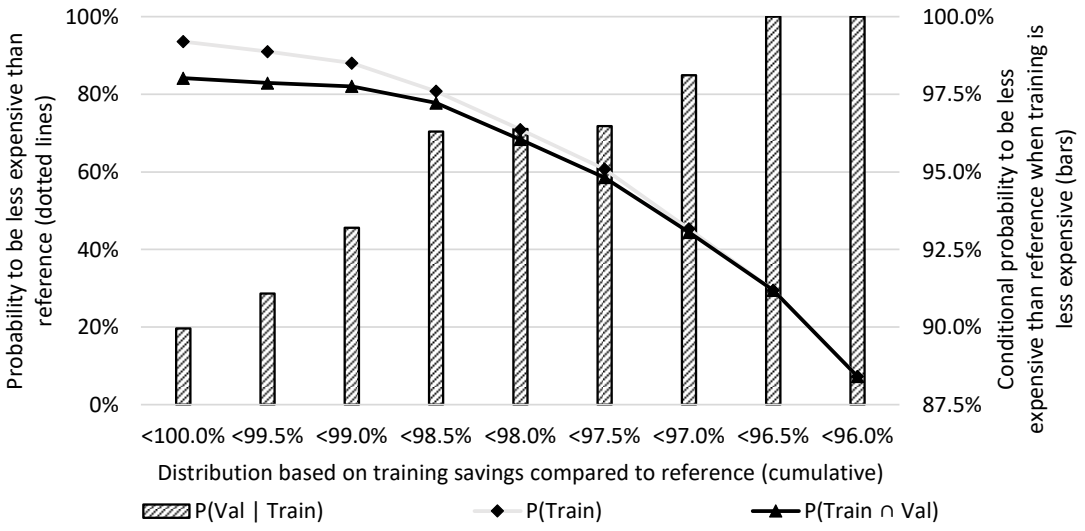

Figure 8. Probabilities of trained agents to reduce operational costs on validation dataset, depending on their training savings.

It can clearly be seen that the most successful agents of the training dataset (on the right side of the Figure 8), with operating cost reductions of more than 4%, have only a small share of the total number of trained agents, but 100% of them can also show an operating cost reduction on the validation dataset.

Finally, a clarification is necessary regarding what average savings can be expected from the agents on the validation dataset depending on their training success (i.e., the operating cost reduction on the training dataset). Figure 9 shows the average savings of the trained agents on the validation dataset, plotted on the same distribution of savings of the trained agents on the training dataset. Thus, agents with higher training success and therefore with higher operating cost savings on the training dataset also may have sharply increased operating cost savings on the validation dataset. Thus, the best agents in training with operating cost savings above 4.1% show an average operating cost saving of 2.8% when applied to the validation dataset. The agent with the highest training success (savings of 4.3%) also has the highest value of all agents on the validation dataset, at 3.3%.

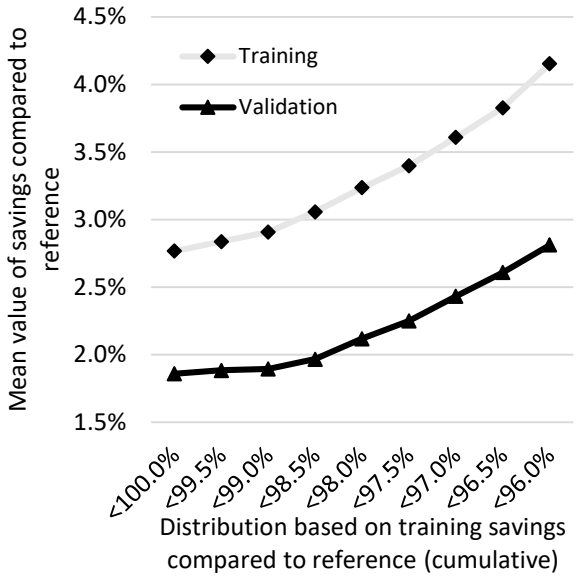

Figure 9. Average savings of trained agents on the validation dataset.

4.2. Operating Cost Comparison

After identifying the best agent based on training success, the influence of its control approach on the operational behavior of the energy system under consideration is now analyzed in more detail. Figure 10 shows representative results of the energy system model controlled with the agent with the highest training success on the validation dataset compared to the reference control.

- In Figure 10a, the operating cost reduction for each time step of the validation dataset is shown. These results can be divided into two areas: one area with only very small deviations of the operating costs of the reinforcement learning control from those of the reference control, and another area with strong operating cost reductions. In addition, a few outliers become obvious in which the reinforcement learning (RL) control seems to take unfavorable decisions (i.e., negative values).
- In Figure 10b, the comparison of the selected flow rates by the RL agent is shown. A strong variation between the available actions or flow rates of the agent can be seen, with the selection of the highest available flow rate mainly in the summer months. The flow rate of the reference control is also shown for comparison.
- In Figure 10c, the COP increase for the domestic hot water supply when RL control is applied. Again, some time steps with unfavorable effects can be seen on the COP of the domestic hot water supply. There, the agent selected an action that leads to a decrease in cooling fluid temperature and therefore a decrease of the COP. A clustering of the COP indicates an increase of about 0.1 in the transition months and 0.3 in the summer months.
- In Figure 10d, the COP increase for the supply of space heating by the RL control is shown. The advantage here is more pronounced compared to the COP of the domestic

hot water supply. Especially in the transition months, strong COP increases can be detected for the space heating supply.

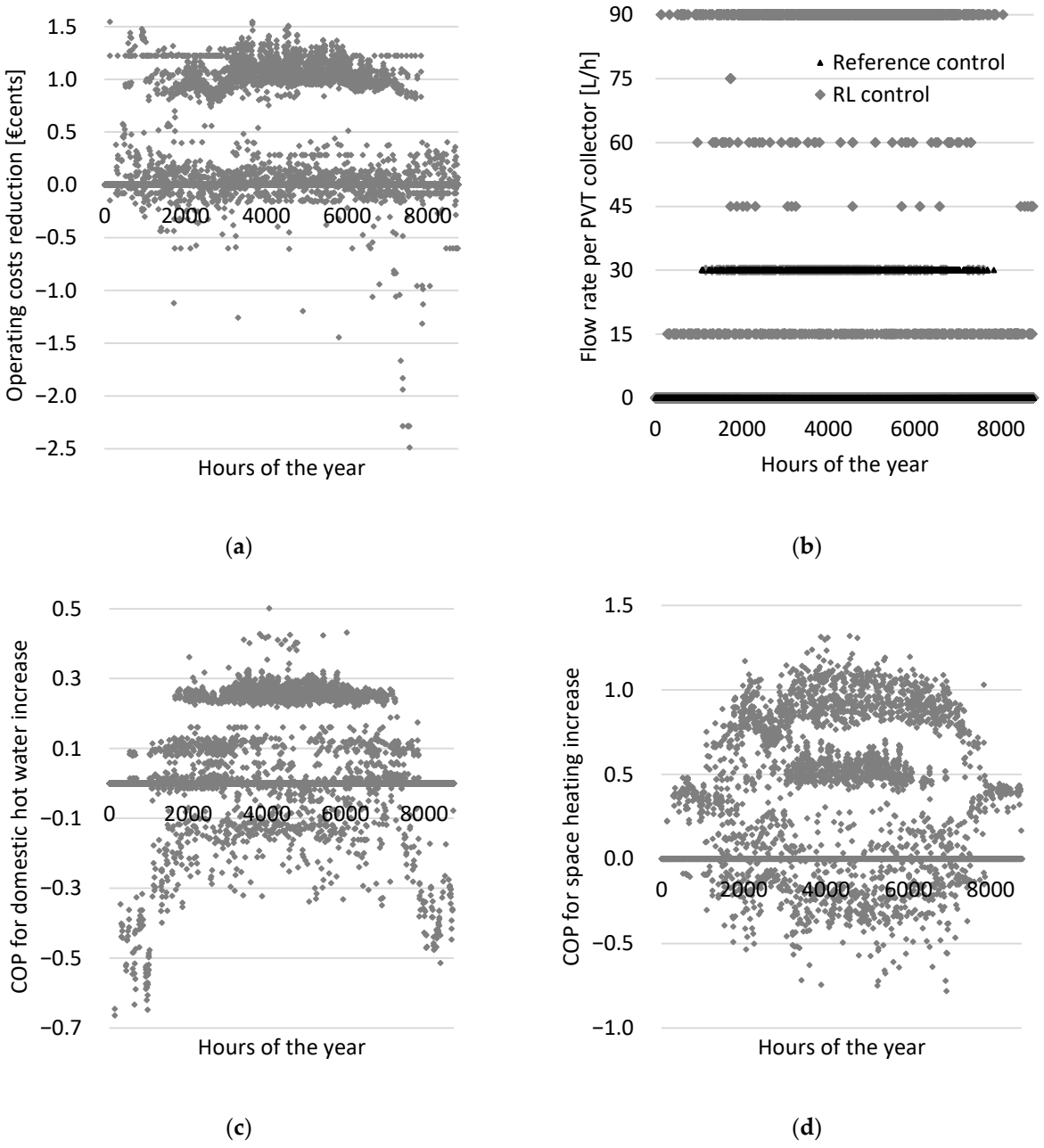

Figure 10. Results of the energy system model with reinforcement learning (RL) control compared to the reference control—Operating cost reduction (**a**), flow rate per PVT collector for reference and RL control (**b**), COP for domestic hot water increase (**c**), and COP for space heating increase (**d**).

Table 4 shows the evaluation of this particular agent with the highest training success on the validation dataset compared to the reference control. For this purpose, the energy quantities for covering the electricity demand as well as the heat pump electricity demand and the self-consumption of the PVT electricity were taken from the model used. Therefore, the biggest difference, caused by the reinforcement learning control, is both the increased self-consumption of the PVT electricity to cover the electrical demand of the simulated single-family house and the decreased electricity demand of the heat pump, caused by the increased coefficient of performance due to the PVT low-temperature heat.

Table 4. Evaluation results of the agent with highest training success compared to the reference control.

	$E_{Grid,Elec.}$ [kWh]	$E_{PVT,Elec.}$ [kWh]	$E_{Grid,HP}$ [kWh]	$E_{PVT,HP}$ [kWh]	$E_{PVT,surplus}$ [kWh]	Self-Consumption [%]
Reference control	3349	1814	1520	374	3567	38.0
RL control	3349	1934	1525	349	3512	39.4
Deviation [%]	-	6.6	0.3	−6.7	−1.5	3.6

Therefore, the operating costs of a complex energy systems, fed by a PVT system, can be reduced by a control approach using reinforcement learning. The development of necessary devices and communication equipment to implement this control in a real single-family house represents the next interesting step in the evaluation of the demonstrated potential.

5. Conclusions

In this paper, the potential of the reinforcement learning approach to control a PVT-heat-pump system by adjusting the flow rate through the included PVT collectors was investigated. In particular, a Bayesian optimization of relevant hyperparameters was used on a training dataset to train an agent for optimal control. The trained agents were then applied to a validation dataset.

Based on the results presented, the following conclusions can be drawn. Conclusions according to the insights of the controller development by applying the reinforcement learning approach are outlined below.

- Bayesian optimization is an applicable approach for selecting promising hyperparameters. Already close to 100 iterations are sufficient to train agents successfully.
- The hyperparameters learning rate, mini-batch size, and discount factor show the strongest influence on the training success of the reinforcement learning approach for the energy system studied here.
- The training success of the reinforcement learning approach (the objective function of Bayesian optimization is used as a measure) can also be repeated on a validation dataset independent of the training dataset, thus demonstrating the generalization capability of the trained agents.

Insights from the operating cost reduction by the developed control approach based on reinforcement learning are summarized below.

- The reinforcement learning approach shows potential as a promising control approach for complex energy systems, illustrated here by a PVT-heat-pump-system. By controlling the flow rate through the PVT system alone, a stable effect in terms of operating cost reduction can be demonstrated.
- The selection of the most successfully trained agents shows a relevant operational cost saving of about 3% on a validation dataset.

Based on the results presented, the question of whether other control variables of a complex energy system based on PVT can also be successfully controlled by reinforcement learning should be investigated. An implementation of the approach presented here in a real single-family house should be developed in future work.

Author Contributions: Conceptualization, D.J.; Formal analysis, D.J.; Investigation, D.J.; Methodology, D.J.; Resources, M.K.; Supervision, M.K.; Validation, D.J.; Visualization, D.J.; Writing—original draft, D.J.; Writing—review and editing, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Projektnummer 491268466 and the Hamburg University of Technology (TUHH) in the funding programme Open Access Publishing.

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

Nomenclature

$C_{Feed-in}$	feed-in tariff
C_{HP}	heat pump electricity tariff
C_{Grid}	electricity price from public grid
$C_{operating}$	operating costs
$E_{Grid,Elec.}$	electricity purchased from the grid for electricity demand
$E_{Grid,HP}$	electricity purchased from the grid for heat pump electricity demand
$E_{PVT,Elec.}$	self-consumption of PVT electricity
$E_{PVT,surplus}$	excess PVT electricity
L	loss function
M	mini-batch size
$\dot{m}_{brine}$	mass flow of brine
$\dot{m}_{PVT}$	mass flow of cooling fluid
T_{amb}	ambient temperature
$T_{heat,supply}$	supply temperature of heating circuit
$T_{heat\ pump,source}$	source temperature for the heat pump
T_{brine}	brine temperature
$T_{PVT,out}$	PVT outlet temperature
Abbreviations	
ANN	artificial neural network
COP	coefficient of performance
DQN	deep Q-network
LSTM	long short-term memory
PV	photovoltaic
PVT	photovoltaic-thermal
RL	reinforcement learning
SFH	single family house

References

1. Abdullah, A.L.; Misha, S.; Tamaldin, N.; Rosli, M.; Sachit, F.A. Theoretical study and indoor experimental validation of performance of the new photovoltaic thermal solar collector (PVT) based water system. *Case Stud. Therm. Eng.* **2020**, *18*, 100595. [CrossRef]
2. Fayaz, H.; Nasrin, R.; Rahim, N.A.; Hasanuzzaman, M. Energy and exergy analysis of the PVT system: Effect of nanofluid flow rate. *Sol. Energy* **2018**, *169*, 217–230. [CrossRef]
3. Srimanickam, B.; Vijayalakshmi, M.M.; Natarajan, E. Energy and Exergy Efficiency of Flat Plate PVT Collector With Forced Convection. *J. Test. Eval.* **2018**, *46*, 783–797. [CrossRef]
4. Wolf, M. Performance analyses of combined heating and photovoltaic power systems for residences. *Energy Convers.* **1976**, *16*, 79–90. [CrossRef]
5. Braunstein, A.; Kornfeld, A. On the Development of the Solar Photovoltaic and Thermal (PVT) Collector. *IEEE Trans. Energy Convers.* **1986**, *EC-1*, 31–33. [CrossRef]
6. Sathe, T.M.; Dhoble, A.S. A review on recent advancements in photovoltaic thermal techniques. *Renew. Sustain. Energy Rev.* **2017**, *76*, 645–672. [CrossRef]
7. Lamnatou, C.; Chemisana, D. Photovoltaic/thermal (PVT) systems: A review with emphasis on environmental issues. *Renew. Energy* **2017**, *105*, 270–287. [CrossRef]
8. Sultan, S.M.; Efzan, M.N.E. Review on recent Photovoltaic/Thermal (PV/T) technology advances and applications. *Sol. Energy* **2018**, *173*, 939–954. [CrossRef]
9. Mustapha, M.; Fudholi, A.; Yen, C.H.; Ruslan, M.H.; Sopian, K. Review on Energy and Exergy Analysis of Air and Water Based Photovoltaic Thermal (PVT) Collector. *Int. J. Power Electron. Drive Syst.* **2018**, *9*, 1367–1373. [CrossRef]
10. Joshi, S.S.; Dhoble, A.S. Photovoltaic -Thermal systems (PVT): Technology review and future trends. *Renew. Sustain. Energy Rev.* **2018**, *92*, 848–882. [CrossRef]
11. Bandaru, S.H.; Becerra, V.; Khanna, S.; Radulovic, J.; Hutchinson, D.; Khusainov, R. A Review of Photovoltaic Thermal (PVT) Technology for Residential Applications: Performance Indicators, Progress, and Opportunities. *Energies* **2021**, *14*, 3853. [CrossRef]
12. Yu, Y.; Long, E.; Chen, X.; Yang, H. Testing and modelling an unglazed photovoltaic thermal collector for application in Sichuan Basin. *Appl. Energy* **2019**, *242*, 931–941. [CrossRef]

13. Herrando, M.; Pantaleo, A.M.; Wang, K.; Markides, C.N. Solar combined cooling, heating and power systems based on hybrid PVT, PV or solar-thermal collectors for building applications. *Renew. Energy* **2019**, *143*, 637–647. [CrossRef]
14. Herrando, M.; Ramos, A.; Zabalza, I.; Markides, C.N. A comprehensive assessment of alternative absorber-exchanger designs for hybrid PVT-water collectors. *Appl. Energy* **2018**, *235*, 1583–1602. [CrossRef]
15. Fudholi, A.; Zohri, M.; Rukman, N.S.B.; Nazri, N.S.; Mustapha, M.; Yen, C.H.; Mohammad, M.; Sopian, K. Exergy and sustainability index of photovoltaic thermal (PVT) air collector: A theoretical and experimental study. *Renew. Sustain. Energy Rev.* **2018**, *100*, 44–51. [CrossRef]
16. Yandri, E. Development and experiment on the performance of polymeric hybrid Photovoltaic Thermal (PVT) collector with halogen solar simulator. *Sol. Energy Mater. Sol. Cells* **2019**, *201*, 110066. [CrossRef]
17. Alous, S.; Kayfeci, M.; Uysal, A. Experimental investigations of using MWCNTs and graphene nanoplatelets water-based nanofluids as coolants in PVT systems. *Appl. Therm. Eng.* **2019**, *162*, 114265. [CrossRef]
18. Hissouf, M.; Feddaoui, M.; Najim, M.; Charef, A. Numerical study of a covered Photovoltaic-Thermal Collector (PVT) enhancement using nanofluids. *Sol. Energy* **2020**, *199*, 115–127. [CrossRef]
19. Madu, K.E.; Uyaelumuo, A.E. Water Based Photovoltaic Thermal (PVT) Collector with Spiral Flow Absorber: An Energy and Exergy Evaluation. *Equat. J. Eng.* **2018**, *2018*, 51–58.
20. Singh, H.P.; Jain, A.; Singh, A.; Arora, S. Influence of absorber plate shape factor and mass flow rate on the performance of the PVT system. *Appl. Therm. Eng.* **2019**, *156*, 692–701. [CrossRef]
21. Hossain, M.S.; Pandey, A.K.; Selvaraj, J.A.; Rahim, N.A.; Islam, M.M.; Tyagi, V.V. Two side serpentine flow based photovoltaic-thermal-phase change materials (PVT-PCM) system: Energy, exergy and economic analysis. *Renew. Energy* **2019**, *136*, 1320–1336. [CrossRef]
22. Barbu, M.; Darie, G.; Siroux, M. A Parametric Study of a Hybrid Photovoltaic Thermal (PVT) System Coupled with a Domestic Hot Water (DHW) Storage Tank. *Energies* **2020**, *13*, 6481. [CrossRef]
23. Senthilraja, S.; Gangadevi, R.; Marimuthu, R.; Baskaran, M. Performance evaluation of water and air based PVT solar collector for hydrogen production application. *Int. J. Hydrogen Energy* **2020**, *45*, 7498–7507. [CrossRef]
24. Christ, D.; Kaltschmitt, M. Modelling of photovoltaic-thermal collectors for the provision of electricity and low temperature heat—Comparison of different flow rate control approaches to optimize the electrical yield. *Renew. Energy Focus* **2021**, *37*, 1–13. [CrossRef]
25. Hengel, F.; Heschl, C.; Inschlag, F.; Klanatsky, P. System efficiency of pvt collector-driven heat pumps. *Int. J. Thermofluids* **2020**, *5–6*, 100034. [CrossRef]
26. Emmi, G.; Zarrella, A.; de Carli, M. A heat pump coupled with photovoltaic thermal hybrid solar collectors: A case study of a multi-source energy system. *Energy Convers. Manag.* **2017**, *151*, 386–399. [CrossRef]
27. Rijvers, L.; Rindt, C.; de Keizer, C. Numerical Analysis of a Residential Energy System That Integrates Hybrid Solar Modules (PVT) with a Heat Pump. *Energies* **2022**, *15*, 96. [CrossRef]
28. Herrando, M.; Ramos, A.; Freeman, J.; Zabalza, I.; Markides, C.N. Technoeconomic modelling and optimisation of solar combined heat and power systems based on flat-box PVT collectors for domestic applications. *Energy Convers. Manag.* **2018**, *175*, 67–85. [CrossRef]
29. Zarei, A.; Liravi, M.; Rabiee, M.B.; Ghodrat, M. A Novel, eco-friendly combined solar cooling and heating system, powered by hybrid Photovoltaic thermal (PVT) collector for domestic application. *Energy Convers. Manag.* **2020**, *222*, 113198. [CrossRef]
30. Vallati, A.; Ocłoń, P.; Colucci, C.; Mauri, L.; de Lieto Vollaro, R.; Taler, J. Energy analysis of a thermal system composed by a heat pump coupled with a PVT solar collector. *Energy* **2019**, *174*, 91–96. [CrossRef]
31. Abu-Rumman, M.; Hamdan, M.; Ayadi, O. Performance enhancement of a photovoltaic thermal (PVT) and ground-source heat pump system. *Geothermics* **2020**, *85*, 101809. [CrossRef]
32. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
33. Yang, L.; Nagy, Z.; Goffin, P.; Schlueter, A. Reinforcement learning for optimal control of low exergy buildings. *Appl. Energy* **2015**, *156*, 577–586. [CrossRef]
34. Raman, N.S.; Devraj, A.M.; Barooah, P.; Meyn, S.P. Reinforcement Learning for Control of Building HVAC Systems. In Proceedings of the 2020 American Control Conference (ACC), Denver, CO, USA, 1–3 July 2020. [CrossRef]
35. Namatēvs, I. Deep Reinforcement Learning on HVAC Control. *Inf. Technol. Manag. Sci.* **2018**, *21*, 29–36. [CrossRef]
36. McKee, E.; Du, Y.; Li, F.; Munk, J.; Johnston, T.; Kurte, K.; Kotevska, O.; Amasyali, K.; Zandi, H. Deep Reinforcement Learning for Residential HVAC Control with Consideration of Human Occupancy. In Proceedings of the 2020 IEEE Power & Energy Society General Meeting (PESGM), Montreal, QC, Canada, 2–6 August 2020. [CrossRef]
37. Ding, X.; Du, W.; Cerpa, A. Octopus: Deep reinforcement learning for holistic smart building control. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, New York, NY, USA, 13–14 November 2019; pp. 326–335. [CrossRef]
38. Zhang, Z.; Zhang, C.; Lam, K.P. A Deep Reinforcement Learning Method for Model-based Optimal Control of HVAC Systems. In *EC-1 Environmental Control Equipment and Systems*; Syracuse University: Syracyse, NY, USA, 2018. [CrossRef]
39. *Matlab*; The MathWorks, Inc.: Natick, MA, USA, 2012.

40. Hoval Thermalia Comfort (6-17), Comfort H (7, 10) (2019). Available online: https://docplayer.org/194149594-Hoval-thermalia-comfort-6-17-comfort-h-7-10-sole-wasser-wasser-wasser-waermepumpe.html (accessed on 30 March 2022).
41. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with Deep Reinforcement Learning. *arXiv* **2013**, arXiv:1312.5602.
42. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2222–2232. Available online: https://arxiv.org/pdf/1503.04069.pdf (accessed on 30 March 2022). [CrossRef]
43. Feurer, M.; Springenberg, J.T.; Hutter, F. *Using Meta-Learning to Initialize Bayesian Optimization of Hyperparameters*; MetaSel@ ECAI: Prague, Czech Republic, 2014. Available online: http://ceur-ws.org/Vol-1201/paper-03.pdf (accessed on 30 March 2022).
44. Hertel, L.; Baldi, P.; Gillen, D.L. Quantity vs. Quality: On Hyperparameter Optimization for Deep Reinforcement Learning. *arXiv* **2020**, arXiv:2007.14604.
45. Stein, B.; Loga, T.; Diefenbach, N. TABULA Web Tool. 2017. Available online: https://webtool.building-typology.eu/#bm (accessed on 1 April 2022).
46. VDI. *Reference Load Profiles of Residential Buildings for Power, Heat and Domestic Hot Water as Well as Reference Generation Profiles for Photovoltaic Plants (VDI 4655)*; Engl. VDI-Gesellschaft Energie und Umwelt: Düsseldorf, Germany, 2021.
47. Pfenninger, S.; Staffell, I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. *Energy* **2016**, *114*, 1251–1265. [CrossRef]
48. Staffell, I.; Pfenninger, S. Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* **2016**, *114*, 1224–1239. [CrossRef]
49. Ruhnau, O. When2Heat Heating Profiles. *Open Power Syst. Data* **2019**. [CrossRef]
50. Tjaden, T.; Bergner, J.; Weniger, J.; Quaschning, V. *Repräsentative Elektrische Lastprofile für EinfamilienhäUser in Deutschland auf 1-Sekündiger Datenbasis*; Hochschule für Technik und Wirtschaft Berlin: Berlin, Germany, 2015.