Intelligent Electric Water Heater Control with Varying State Information

Christophe Patyn^{*†‡}, Thijs Peirelinck^{*‡}, Geert Deconinck^{*}, Ann Nowé[†]

^{*}Department of Electrical Engineering, KU Leuven/EnergyVille, 3001 Leuven, Belgium

[†]Department of Computer Science, Vrije Universiteit Brussel, 1050 Brussel, Belgium

[‡]These authors contributed equally to this work

Abstract—The increasing share of renewable energy sources in the electricity grid results in a higher degree of uncertainty regarding electrical energy production. In response to this, flexibility of the demand has been proposed as part of the solution. An important source of flexibility available at the residential consumer side are thermostatically controlled loads (TCLs). In this paper the activation of this source of flexibility is achieved by applying batch reinforcement learning (BRL) to an electric water heater (EWH) in a Time of Use (ToU) setting. The cost performance of six BRL agents with six different state spaces is compared quantitatively. In every case, the BRL agent can successfully shift energy consumption within 20-25 days. The performance of an agent with access to multiple temperature sensors along the height of the EWH is comparable to the performance of an agent with access to only the highest temperature sensor. This indicates manufacturing costs related to sensors can be reduced while maintaining the same performance. Additionally, results show that the inclusion of a theoretical state of charge value in the state space increases performance by more than 8% compared to the performance of the other BRL agents. It is therefore argued that an estimation of the state of charge should be included in future work as it would increase cost performance.

I. INTRODUCTION AND LITERATURE REVIEW

The importance of active grid management rises as the share of intermittent renewable energy resources in the energy mix increases. With more renewable energy sources, such as wind or solar generation, comes more uncertainty, which dignifies active grid management compared to the fit-and-forget approach often applied in the past. Demand Response (DR) is one of the techniques leveraged to activate the full capacity of the grid, including its peripheral loads. It includes all intentional modifications to consumption patterns of electricity of end use customers that are intended to alter the timing, level of instantaneous demand, or the total electricity consumption [1]. DR makes use of flexibility carriers, such as Thermostatically Controlled Loads (TCLs), electric vehicles, etc., to enable this consumption modification. When a larger portion of the demand for electricity can be matched with periods of high renewable energy generation, the possibility of running the electric grid with a higher penetration of solar, wind, etc. comes ever closer.

DR control is a sequential decision-making problem which results in a Partially Observable Markov Decision Process

(POMDP) with stochastic environment dynamics. The partial observability stems from the fact that certain state variables cannot be measured exactly, such as the thermal energy contained in a building envelope or in an Electric Water Heater (EWH). Model Predictive Control (MPC) [2] was proposed multiple times in this regard. For example, Kalman filters are often used in MPC to estimate hidden state features such as the temperature of a building envelope [3] or the State of Charge (SoC) of an EWH [4]. Other work proposed to take the heterogeneity of a cluster of TCLs into account by modeling tracer devices [5]. Reinforcement learning [6] algorithms have been applied in recent research to capture the environment dynamics and solve the POMDP completely model-free by only relying on sensory inputs. Specifically Batch Reinforcement Learning (BRL) has proven successful in the residential DR context [7], [8]. The BRL algorithm Fitted Q-Iteration (FQI) is implemented in [9] to control a cluster of EWHs where the aggregator receives an average water temperature from each EWH. In [10], the same BRL algorithm is applied to another TCL. At training time, the exogenous components of the state vector are replaced with its forecasts and a model-free Monte Carlo method is implemented to find a consumption plan. The proposed extension outperforms regular BRL after 20-25 days. The authors of [11] apply model-assisted BRL for building climate control subjected to dynamic pricing. They approximate the environments transition function with extreme learning machines to obtain virtual transition tuples and propose policy shaping with triangular membership functions. However, both additions were shown to have a limited positive impact on performance in both simulation and living lab environments. De Somer et al. implemented a similar model-based BRL approach for the control of a Domestic Hot Water (DHW) buffer in [12]. The controller aims to maximize self-consumption of the rooftop PV installation. To minimize interaction with the EWH the authors use a model of the buffer as well as the DHW demand to generate training samples. Moreover, to simplify the temperature distribution along the height of the buffer only complete heating cycles are allowed in their approach. The authors show that their active control strategy almost triples the PV production captured by heating DHW, compared to regular thermostat control, in a real life implementation on six houses. Claessens et al. control TCLs in a district heating network and obtain a performance within 65% of a theoretical lower bound on the

This work was supported in part by VLAIO-SBO project 140047: SMILE-IT.

cost. However, they propose a handcrafted feature extraction method to evaluate what performance can be obtained from this limited state description [13]. More sophisticated feature extraction techniques are proposed by Claessens et al. in [14], where a binning of the state space of a cluster of TCLs allows for a convolutional neural network to be used. In [15], similar deep learning techniques such as convolutional and recurrent neural networks were proposed to approximate near-optimal policies for a residential heating system and an EWH. The authors show results where the recurrent network outperforms other tested models after about 15-20 days.

In this paper, a Reinforcement Learning (RL) agent controls an EWH with the FQI algorithm. The agent is incentivized by a day-ahead market price to shift the load to low-price periods. The algorithm only uses sensory data. The performance of this implementation is therefore strongly dependent on the available sensors. If less sensors are needed, manufacturing costs will decrease. Furthermore, the data collection will be simplified making the implementation of the controller easier in a real life setting. Therefore, the cost performance of different BRL agents with different sensory inputs is compared here. The contribution of this paper is that the best cost performance can be attained with the *SoC* of the EWH. As this is an unobservable variable, it is argued that future work should focus on estimating its value.

The remainder of this paper formulates the considered control problem as a Markov Decision Process (MDP) in section II. With this formulation in mind, section III introduces the algorithm used to solve the sequential decision-making problem. Section IV then elaborates on the TCL and its simulation model considered here. Next, section V shows the results. Section VI concludes this paper.

II. MARKOV DECISION PROCESS

Following the procedure presented in [10], an agent is tasked with the control of an EWH at minimal cost. DHW always needs to be available when the user requests it. To comply with this constraint, the outlet water temperature T_{out} of the EWH is required to stay above a minimum temperature T_{min} . DHW consumption data is used to simulate user behaviour. The agent's actions are subject to a variable Belpex day-ahead market price for electricity. To arrive at a solution that shifts the EWH's electric load to low-price periods, a cost-minimizing objective is formalized. The minimization horizon T equals 96 time steps t , or one day, with Δt equal to 15 minutes. As is usual in reinforcement learning, the problem is modelled in the following paragraphs as a MDP.

A. State Space X

Based on the environment's current state, the agent decides which action to execute. The observed state $x^{\text{obs}} \in X^O$ only includes the values of the state features at time step t . Not all state information is directly observable. The observed state-space is thus a subset of the full state space of the EWH: $X^O \subset X$. Therefore, this problem is actually a POMDP and the Markov property does not hold. This means the current

observed state x^{obs} is not only dependent on the previous state but also on earlier states. As proposed in [16], adding a history h of x^{obs} will allow the agent to recover the Markov property. The resulting augmented state $x^{\text{aug}} \in X^{\text{aug}}$ is described in equation 1.

$$x^{\text{aug}} = [t, [x_a^t, x_a^{t-1}, \dots, x_a^{t-h}], \dots, [x_z^t, x_z^{t-1}, \dots, x_z^{t-h}]] \quad (1)$$

where $x_a^t, \dots, x_z^t$ are the features assumed to be observable by the agent and $t \in \{1, \dots, 96\}$ is the time step.

The partial observability affect the agent's perception of the environment's transitions. This can be described by equation 2, in which f_T is the stochastic discrete-time transition function. The current augmented state x^{aug} of the environment and the chosen action u lead to the next state $x^{\text{aug}'}$.

$$x^{\text{aug}'} = f_T(x^{\text{aug}}, u) \quad (2)$$

This paper aims to compare the performance of different state spaces X^O for the control of the EWH. Table I describes the different configurations for x^{obs} that were tested. Configuration $X1$ includes the top layer water temperature T_{out} and the DHW demand m in addition to the current time step t , which is related to the Time-of-Use (ToU) price profile. The hypothesis of the authors is that T_{out} can indicate both the *SoC* and its lower boundary, as it is the only input to the backup controller (section IV-B). Additionally, combining m with T_{out} should give the agent an indication of the energy outflow. $X2$ adds the consumed electric energy ΔE during that time step, which serves as a proxy for the amount of added thermal energy. $X3$ includes the water temperature T_{supply} at the bottom of the EWH as well. This temperature value in combination with T_{out} might enable the agent to more accurately estimate the *SoC*. The fourth configuration takes this a step further by incorporating another three temperature sensors T_1 , T_2 and T_3 along the height of the buffer. The choice for these particular state features in $X1$ to $X4$ is motivated by a desire to understand the minimal information needed to operate the EWH. A minimal amount of information leads to a minimum amount of sensors that are needed to provide the necessary data, which reduces manufacturing costs. Two extra configurations have been tested to push this idea to the limit. $X5$ only considers the consumed electric energy ΔE , while $X6$ adds the *SoC* of the EWH in addition to the time. The *SoC* is defined in section IV-A and provides a measure for the available energy content in the buffer, relative to the maximal energy content possible. This is however an unobservable variable which would have to be estimated in a real life setting. The results will indicate the size of the performance increase if such an estimate would be obtained. In this paper, the *SoC* is considered to be given.

B. Action Space U

At each time step t , the agent takes a binary control action $u \in U = \{0, 1\}$ in order to turn the heating element of the EWH on or off. In the physical world this corresponds to drawing 0 or 2.4 kW of electric power during a period $\Delta t =$

TABLE I
STATE SPACE CONFIGURATIONS

Symbol	x^{obs}
$X1$	t, T_{out}, m
$X2$	$t, T_{out}, m, \Delta E$
$X3$	$t, T_{out}, T_{in}, m, \Delta E$
$X4$	$t, T_{out}, T_1, T_2, T_3, T_{in}, m, \Delta E$
$X5$	$t, \Delta E$
$X6$	t, SoC

15 minutes, assuming the EWH has a rated power of $P_r = 2.4$ kW.

C. Cost Function Q

In this DR setting, a deterministic ToU price signal λ is available at the start of the optimization horizon T . λ is the Belpex day-ahead market price for electricity. When the agent executes an action u , the electric energy that is consumed is subject to λ . As a MDP is a sequential decision making problem, the agent incurs a cost c at every time step. When these costs are discounted and added over the agent's optimization horizon T , the state-action value function, or the Q -function, is obtained as in equation 3, with γ the discount factor.

$$Q_t(x^{aug}, u_t) = \sum_{i=t}^T \gamma^{i-t} c(u_i) \quad (3)$$

with

$$c(u) = u\lambda P_r \Delta t \quad (4)$$

The goal of the agent is to approximate this Q -function and take cost-minimizing actions at runtime, by solving equation 5.

$$u_t = \operatorname{argmin}_u Q_t(x^{aug}, u) \quad (5)$$

III. FITTED Q-ITERATION

The Q -function provides the long term cost of a state-action pair and needs to be approximated by the agent. In this work, the FQI algorithm is implemented for this purpose. Algorithm 1 shows the pseudocode. $x^{aug'}$ defines the next state of the environment after applying u . At execution time, the agent collects transition tuples $(x^{aug}, u, x^{aug'}, \Delta E)$, which are sampled by executing actions in the environment, in a data batch $\mathcal{F}$. At training time, Q -value targets are recomputed for all tuples k in $\mathcal{F}$. First, the immediate cost of an action is calculated with new ToU price λ and the consumed energy ΔE of tuple k (line 5). In a second step, the new Q -value that corresponds to tuple k is found according to the Bellman equation (line 6). All state-action pairs with their new Q -values are stored in the training set $\mathcal{TS}$. The Q -function $Q(x^{aug}, u)$ is then approximated by a neural network model $\hat{Q}_t$ according to a receding horizon approach [9] using the data in $\mathcal{TS}$ (line 3).

Algorithm 1 Fitted Q-iteration

```

1: Inputs:  $\mathcal{F} = \{(x^{aug}, u, x^{aug'}, \Delta E)_k\}_{k=1}^{\#\mathcal{F}}, \lambda$ 
2:  $\forall x_k^{aug} \in X^{aug}, u_k \in U : \hat{Q}_T = 0$ 
3: for  $t = T - 1, \dots, 1$  do
4:   for  $k = 1, \dots, \#\mathcal{F}$  do
5:      $c_k \leftarrow \lambda_t \Delta E_k$ 
6:      $Q_k \leftarrow c_k + \min_{u \in U} \hat{Q}_{t+1}(x^{aug'}, u)$ 
7:   end for
8:   use approximator to obtain  $\hat{Q}_t$  from
      $\mathcal{TS} = \{(x_k^{aug}, u_k, Q_k), k = 1, \dots, \#\mathcal{F}\}$ 
9: end for
10: Output:  $[\hat{Q}_1, \dots, \hat{Q}_T]$ 

```

IV. ENVIRONMENT

The TCL serving as the environment for the RL agent is a single EWH with buffer. In order to simulate the impact of the agent's decisions on the dynamics of this EWH a thermally stratified model of the buffer has been used. The model considered is based on existing literature [17] and has been validated in a lab [18]. Apart from the thermal model of the tank, also user comfort and behavior has been simulated. All these different components of the environment have been explained next. The buffer model is extensively described in [9] and [18]. Therefore, only the parts relevant for this work have been presented.

A. EWH simulation model

A DHW buffer model based on recent design standards has been considered. Hot water leaves the tank at the top, and is replaced by cold water at the bottom. A resistance, located at the bottom, can inject heat in the tank. During simulations, requested hot water temperature T_{asked} is assumed constant and equal to T_{min} . As a result, to ensure comfort, the output water temperature T_{out} is never allowed to be below T_{min} . Also inlet cold water temperature T_{supply} has been assumed constant. However, T_{supply} can be lower than T_{min} . Moreover, it has been assumed the boiler cannot influence its environment's temperature T_{env} , and that this temperature is also constant. Furthermore, the EWH has a rated power P_r and buffer volume V . The values of these parameters and other important model and simulation parameters have been summarized in table II.

The water in the DHW buffer is divided in n layers. Each layer i has a uniform temperature T_i , and volume $V_i = V/n$. Since $T_{supply} < T_{min}$ during simulation, several layers can have a temperature $T_i < T_{min}$. Thermal losses, conduction effects and mixing are defined as in [9], with the parameters summarized in table II. In different simulation scenarios, different sensors along the height of the buffer have been assumed to be available. The different scenarios and their assumption on available sensors have been discussed in section II.

The EWH-model has been simulated in discrete time-steps of length Δt_s . At each simulation step all layer temperatures

TABLE II
PARAMETER VALUES

Simulation parameter	Value	Unit
T_{env}	20	°C
T_{supply}	10	°C
T_{asked}	45	°C
T_{max}	60	°C
V	160	l
P_r	2.4	kW
Model parameter	Value	Unit
n	50	-
Δt_s	5	s
C (water heat capacity)	4190	J/kg K
k (water thermal conductivity)	0.59	W/mK

are calculated. The water of the lowest layer can be heated by setting the resistance at the bottom of the tank on. The resistance is on until heating is stopped or until all layers reach T_{max} . Vanthournout et al. explain all details of the heating process in [18].

State space configuration X6 assumes a *SoC* measurement of the energy content in the buffer is available. In the model, *SoC* is defined as in equation 6 [18]. Where $L = \{i \in \{1, \dots, n\} | T_i \geq T_{\min}\}$. The *SoC* is thus equal to the ratio of the energy content of the layers satisfying the $T_i \geq T_{\min}$ comfort constraint, versus the maximal possible energy content given the $T_i \leq T_{\max}$ constraint. Clearly, it is not realistic nor cost-effective to measure the temperature along the entire height of the buffer, for each layer i . State configuration X6 is therefore only interesting when it is possible to estimate the *SoC* of the EWH from the provided data.

$$SoC = \frac{\sum_{l \in L} V_l (T_l - T_{\min})}{V (T_{\max} - T_{\min})} \quad (6)$$

B. User comfort

As mentioned earlier, the outlet water temperature T_{out} has to be above the minimal comfortable temperature $T_{\min}$, to take into account comfort constraints of the users. To ensure this, the EWH is equipped with a backup controller. This backup controller has the power to overrule the action of the RL agent. It will do so by turning the EWH on (off) when the temperature is too cold (hot), regardless of the agent's action.

The backup controller defines the actual physical action u_{phys} of the EWH by

$$u_{\text{phys}} = \begin{cases} 1 & T_b \leq T_{\min} \\ u & T_b > T_{\min} \text{ and } T_{\text{out}} < T_{\max} \\ 0 & T_{\text{out}} \geq T_{\max} \end{cases} \quad (7)$$

where T_b is the temperature at the backup sensor location, near the top of the buffer, and u is the action given by the agent.

C. User behavior

At any given moment in time a user can request a certain amount of hot water from the buffer. As mentioned earlier, the

Fig. 1. Example of boiler operation for one day when only the backup controller is active. The top plot depicts T_i of 50 equidistant water layers i in the buffer. The top layer is colored dark red, which fades into blue the closer the water layer i is to the bottom of the buffer. The middle plot shows the DHW consumption and the bottom plot the EWH's cumulative energy consumption.

requested temperature T_{asked} is assumed constant and equal to $T_{\min}$. However, outlet temperature can be higher. If this is the case, hot water is mixed with enough cold water to reach T_{asked} .

The Integrated District Energy Assessment Simulations (IDEAS) library [19] provides a tool to generate realistic DHW consumption profiles, based on real consumption data from Belgian households [20]. Here, a profile of a dual-earner family with two children has been used.

Figure 1 shows a one day simulation of the environment. During this day, only the backup controller is active, the agent's action is thus always $u = 0$. Thus, water inside the DHW buffer is only heated when necessary to ensure user comfort. The top figure shows the temperature of all n layers, the middle figure the amount of DHW consumed every quarter, in liters. The bottom figure shows the energy consumed by the EWH. The influence of DHW demand on the temperature profile of the water inside the buffer, and the energy consumption as a result of heating this water is clearly visible, especially at hour 18. At this time, a significant amount of hot water is drawn compared to the total volume of the buffer. This triggers the controller to start heating, resulting in temperature rise and energy consumption. The drainage of hot water creates a challenging optimization problem. The moment and amount of hot water tapped in the future is unknown and therefore the amount of energy contained in the buffer can change abruptly and at any time. Possibly, this deems all previous efforts to avoid an electricity price peak ineffective.

Fig. 2. Cumulative cost for 30-day simulations with 6 different sensory inputs.

V. RESULTS

As mentioned earlier, this paper aims to contribute to the application of BRL in residential DR by applying FQI to an EWH control setting. Particularly of interest is the performance of the FQI algorithm when different sensory inputs are available. The goal of the agent is to minimize operation costs, which are subject to a ToU price profile, while taking into account user comfort. Earlier research either used a model-based control strategy or assumed intrusive sensors were available along the height of the tank. By only considering readily available sensors, combined with model-free learning, the environment of the agent is significantly more challenging.

In this paper, six different state configurations have been tested, which can be reviewed in table I. For each of these configurations, a simulation of thirty days was run ten times. During the first ten days of each simulation, the agent's policy was executed ϵ -greedy [6]. This means that every time the agent has to choose an action, a random number between 0 and 1 is chosen. If the random number is smaller than ϵ , a random action is chosen. Otherwise, the agent chooses the best action. At the start of each simulation, ϵ was initialized to 1, meaning a completely random policy was executed on the first day. ϵ was annealed every day. After day ten, ϵ equals zero and thus the agent's policy is executed fully greedily according to equation 5.

Figure 2 shows the average cumulative cost for the different simulations as a solid line and their 68% confidence interval. The figure also shows the cumulative cost without control, i.e. only the backup controller is active. From the results with intelligent control, three groups can be distinguished. X3 clearly performs worst, indicating that T_{supply} clutters the agent's perception of the current state. This cluttering occurs when the bottom layer has a low temperature while most other layers' temperature is still relatively high. The second group consists of X1, X2, X4 and X5. The more days into the simulation, the more the cumulative cost of these state configurations converge towards each other. Interestingly, this

result indicates the agent can extract a comparable amount of information from one sensor at the exit valve, which measures T_{out} as in X1, compared to five temperature sensors along the height of the buffer as in X4. Since the backup controller only overrides the agent's chosen action u if $T_b \approx T_{\text{out}}$ is under or above T_{min} or T_{max} , it is conceivable that the agent is able to operate the EWH with only the most important information. Furthermore, measuring the consumed electric energy and only using this feature, as in X5, suffices to achieve the same level of performance. The agent can distill the amount of extracted energy from the combination of the T_{out} and the DHW demand m , which is again comparable information to the consumed energy ΔE .

The most efficient state configuration proves to be the one including the SoC . This feature describes the total amount of energy available in the buffer, relative to the maximal amount available given the temperature settings. Intuitively, it is not surprising that, while controlling an energy buffer, SoC provides the most important state information. In reality, determining the SoC of an EWH directly might not be feasible. However, if this value could be estimated it is clear this feature can provide superior information to the agent. On average, after thirty days, X6 is 8.32% cheaper than all other state configurations, and 16.55% cheaper than the scenario where only the backup controller is active.

Figure 3 and 4 visualize the resulting control policies during the same six days for both the worst and best performing state configurations, X3 and X6 respectively. The top part shows the temperature profile of 5 layers, equidistantly chosen along the height of the buffer. The horizontal line indicates the temperature at which the EWH is activated by the backup controller. This is when $T_b = T_{\text{min}} = 45^\circ\text{C}$. The middle figure shows the amount of DHW consumed on a quarter hourly basis. In the bottom part of the charts, the power usage of the EWH is plotted in blue, with the price in grey. It is obvious that both agents do not always manage to evade high prices. This is due to unexpected DHW consumption. And although X6 performs considerably better on cost, this is hard to validate qualitatively by looking at these charts. The differences in policies are very subtle. For example on day 24, the agent with state X6 manages to delay drawing power until prices are low again at the end of the day, while the other agent already turns the EWH on at an earlier time when prices are still relatively high.

VI. CONCLUSION

This paper discussed the control of a TCL, an EWH in particular, by a BRL agent. Its objective is to minimize daily electricity cost while its actions are subject to the day-ahead market price for electricity. Realistic DHW consumption profiles have been obtained and used in the simulations. This to ensure a realistic environment in which the agent has to take user behaviour into consideration.

The problem has been modelled as a MDP, in which the state the agent observes from the environment is expanded with a history of previous states to recover the Markov

Fig. 3. 6 days of X6 operation. The top plot depicts T_i of 5 equidistant water layers i in the buffer, the middle plot shows the DHW consumption and the bottom plot the ToU price (gray) and EWH power consumption (blue)

Fig. 4. 6 days of X3 operation. The top plot depicts T_i of 5 equidistant water layers i in the buffer, the middle plot shows the DHW consumption and the bottom plot the ToU price (gray) and EWH power consumption (blue)

property. The FQI algorithm is implemented to solve this MDP with a receding horizon approach.

The main goal of this paper has been to compare cost performance of agents with varying access to sensory inputs. Six different configurations were defined, with ten simulation runs each. The results show that both the configuration with only the energy consumption and the one with access to the temperature of the top layer of water and the DHW consumption performed equally well as agents with access to many more sensory inputs, which shows the strength of model-free approaches. Furthermore, simulations where the agent only had access to the SoC of the EWH proved to be more than 8% more cost-efficient. Although the SoC is a theoretical value, an estimation of this variable can provide a TCL-agent with valuable information, resulting in superior performance. Future work will therefore focus on using machine learning to

estimate the SoC of an EWH. This estimation technique can then be used as an input to a BRL controller for real world systems.

REFERENCES

- [1] M. H. Albadi and E. F. El-Saadany, "Demand Response in Electricity Markets: An Overview," in *2007 IEEE Power Engineering Society General Meeting*, 2007.
- [2] Y. Ma, "Model Predictive Control for Energy Efficient Buildings," Ph.D. dissertation, UC Berkeley, 2012. [Online]. Available: <https://escholarship.org/uc/item/1dj908bs>
- [3] E. Vrettos, E. C. Kara, J. MacDonald, G. Andersson, and D. S. Callaway, "Experimental Demonstration of Frequency Regulation by Commercial Buildings amp;#x2014;Part I: Modeling and Hierarchical Control Design," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3213–3223, Jul. 2018.
- [4] H. Kazmi, S. D'Oca, C. Delmastro, S. Lodeweyckx, and S. P. Corgnati, "Generalizable occupant-driven optimization model for domestic hot water production in NZEB," *Applied Energy*, vol. 175, pp. 1–15, Aug. 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0306261916305682>
- [5] S. Iacovella, F. Ruelens, P. Vingerhoets, B. Claessens, and G. Deconinck, "Cluster Control of Heterogeneous Thermostatically Controlled Loads Using Tracer Devices," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 528–536, Mar. 2017.
- [6] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2017.
- [7] C. Patyn, F. Ruelens, and G. Deconinck, "Comparing neural architectures for demand response through model-free reinforcement learning for heat pump control," in *2018 IEEE International Energy Conference (ENERGYCON)*, 2018.
- [8] T. Peirelinck, F. Ruelens, and G. Deconinck, "Using reinforcement learning for optimizing heat pump control in a building model in Modelica," in *2018 IEEE International Energy Conference (ENERGYCON)*, 2018.
- [9] F. Ruelens, B. J. Claessens, S. Vandael, S. Iacovella, P. Vingerhoets, and R. Belmans, "Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning," in *Power Systems Computation Conference*, 2014.
- [10] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, 2016.
- [11] G. T. Costanzo, S. Iacovella, F. Ruelens, T. Leurs, and B. Claessens, "Experimental analysis of data-driven control for a building heating system," *Sustainable Energy, Grids and Networks*, 2016.
- [12] O. De Somer, A. Soares, T. Kuijpers, K. Vossen, K. Vanthournout, and F. Spiessens, "Using Reinforcement Learning for Demand Response of Domestic Hot Water Buffers: a Real-Life Demonstration," *IEEE ISGT-Europe*, 2017.
- [13] B. J. Claessens, D. Vanhoudt, J. Desmedt, and F. Ruelens, "Model-Free Control of Thermostatically Controlled Loads Connected to a District Heating Network," *Energy and Buildings*, 2018.
- [14] B. J. Claessens, P. Vrancx, and F. Ruelens, "Convolutional Neural Networks For Automatic State-Time Feature Extraction in Reinforcement Learning Applied to Residential Load Control," *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–1, 2016.
- [15] F. Ruelens, B. J. Claessens, P. Vrancx, F. Spiessens, and G. Deconinck, "Direct Load Control of Thermostatically Controlled Loads Based on Sparse Observations Using Deep Reinforcement Learning," *arXiv preprint arXiv:1707.08553*, 2017.
- [16] M. Hausknecht and P. Stone, "Deep Recurrent Q-learning For Partially Observable MDPs," *Computing Research Repository*, 2015.
- [17] TRNSYS, "TRNSYS : Transient System Simulation Tool." [Online]. Available: <http://www.trnsys.com/>
- [18] K. Vanthournout, R. D'hulst, D. Geysen, and G. Jacobs, "A Smart Domestic Hot Water Buffer," *IEEE Transactions on Smart Grid*, 2012.
- [19] R. Baetens, R. De Coninck, F. Jorissen, D. Picard, L. Helsen, and D. Saelens, "Openideas-an open framework for integrated district energy simulations," in *Building simulation*, 2015.
- [20] D. Aerts, J. Minnen, I. Glorieux, I. Wouters, and F. Descamps, "A probabilistic activity model to include realistic occupant behaviour in building simulations," in *IBPSA-Canada eSim*, 2014.