Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Electric Water Heaters Management via Reinforcement Learning with Time-Delay in Isolated Microgrids

JIANGJIAO XU[1], HISHAM MAHMOOD[2], HAO XIAO[3], ENRICO ANDERLINI[4] AND MOHAMMAD ABUSARA[1]

[1]College of Engineering, Mathematics and Physical Sciences, Exeter University, UK (e-mail: j.xu@exeter.ac.uk)
[2]Department of Electrical and Computer Engineering, Florida Polytechnic University, USA(e-mail: author@lamar.colostate.edu)
[3]Institute of Electrical Engineering (IEE), Chinese Academy of Sciences (CAS), China
[4]Department of Mechanical Engineering, University College London, UK

Corresponding author: Mohammad Abusara (e-mail: m.abusara@exeter.ac.uk).

ABSTRACT Isolated microgrids powered by renewable energy sources, battery storage, and backup diesel generators need appropriate demand response to utilize available energy and reduce diesel consumption efficiently. However, real-time demand-side management has become a significant challenge due to the communication time-delay issue. In this paper, a distributed model-free strategy is proposed to manage the demand of Electric Water Heater (EWH) units. The distributed artificial intelligence technology based on Reinforcement Learning (RL) is adopted to independently control the 150 EWHs using a virtual tariff. Two different strategies are proposed to generate the virtual tariff and they are compared to each other to investigate the impact of communication time-delay to the proposed RL algorithm in real-time control scenario. The first strategy is based on measuring the battery State of Charge (SOC) in real time while the second method is based on predicting the SOC 24-hours in advance using an Artificial Neural Network (ANN). The results show that the communication time-delay greatly influences the convergence result of the first method while the second method showed high immunity. The results also show that the proposed algorithm reduces the use of energy consumption by an average of 8.91%(6.675kW) for each EWH, which symbolizes the viability of the proposed approach.

INDEX TERMS Energy storage, distributed control, reinforcement learning, electric water heaters, Q-learning, time-delay.

Nomenclature

α	The learning rate
η	The learning rate coefficient
μ	The mean value
ρ	The mass density of the water
σ	The standard deviation value
$\varphi_1(t)$	The Gaussian density
$\varphi_2(t)$	The stochastic delay density
A	The cross-sectional vector area
a_t	The action
E_t	Reward for running the EWH
$erf(\cdot)$	The error function
L_t	Reward for water tank temperature
M_f	The mass of water in the full tank
M_t	The demand for inlet cold water to the tank
r_t	Total reward function
s_t	The state
t	The time index
T_o	The ambient temperature ($^\circ C$)
T_t	The current average water temperature in the tank
$Tariff_k$	Virtual tariff
$Temp_l$	Water temperature
ToD	Time of day
v	The flow velocity of the mass elements
Q	The heat rate (kW) of the EWH
UA	The heat loss coefficient

I. INTRODUCTION

THE world is rapidly turning into a global village, and the requirement for energy and other related services is also increasing. However, 1.4 billion people worldwide still lack access to electricity, and about 85% of them are live

in rural areas [1]. The CO_2 emissions from the electricity and commercial heat used in buildings have increased to $10GtCO_2$ [2]. Due to the depletion of fossil fuels and their associated environmental impact, more distributed generators based on Renewable Energy Sources (RES) are penetrating the current power systems market. This will not only mitigate the global climate change caused by fossil fuel but also support social and economic development of remote and isolated communities [3]–[5].

Energy storage is considered as an essential element to balance the generation and demand. Energy management of storage and non-critical loads is also vital to improve the economic performance and reliability of an environmentally friendly power system [6]. The domestic hot water consumption accounts for up to 40% of the total domestic energy usage [7]. An effective control strategy needs to optimize the total power consumption including domestic hot water consumption among renewable generators, energy storage systems, and other facilities to minimize fuel consumption while meeting load demand. It can not only help these traditional power networks upgrade to smart grids, but also reduce the cost of fossil fuels in the entire island power system, optimize the energy structure, and reduce greenhouse gas emissions. Many control and optimization approaches have been investigated to achieve optimal results in energy systems such as Linear Programming (LP) [8], Mixed Integer Linear Programming (MILP) [9], [10], Mixed Integer Non-Linear Programming (MINLP) [11], and Genetic Algorithm (GA) [12]. These existing traditional analytical approaches are quite cumbersome and need several simplifying assumptions. They all require a detailed mathematical model of the system and some of them require system linearisation.

Artificial Intelligence (Al) based methods can, however, perform complex non-linear non-convex optimization and predict the energy demand and generation without the need for a mathematical model [13]. There has been a growing interest in the application of Al-based algorithms in energy systems. Several studies have also been presented to predict power consumption in energy systems, including Artificial Neural Networks (ANN) [14], Multiple Linear Regression (MLR) [15], Support Vector Machine (SVM) [16], and Decision Tree (DT) [17]. For energy prediction in buildings, Mechaqrane et al. [18] presented a performance comparison between a linear Auto-Regressive model with exogenous input (ARX) and a neural network ARX (NNARX) model to forecast the indoor temperature, with the latter resulting in improved efficiency.

In recent years, RL has been used to implement Demand Response (DR) and distributed energy management strategies for smart homes and smart grids [19]–[21]. Xu et al. proposed a completely distributed multi-agent associated with RL to optimize the reactive power dispatch. The proposed Q-learning algorithm can increase the learning speed and achieve near-optimal solutions [22]. In [23], a cooperative RL algorithm is proposed for distributed economic dispatch without using a specific mathematical model. A Markov decision process (MDP) modelled the energy trading process and an RL algorithm was utilized to optimize the decision in the MDP [24]. The simulation results verified the performance of the proposed demand side management system.When interacting with a specific environment, RL-based optimization algorithms can learn and choose actions based on experience [25]. In contrast, traditional optimization methods need specific system's and environmental mathematical models, which require a high degree of data, knowledge of control, and expertise. In [26], a distributed energy management strategy for a combined heat and power system, and a vanadium redox battery was introduced to optimize the discharging policy using RL. A deep RL based energy trading scheme with multiple Microgrids was proposed in [27] to optimize the energy trading policy. Reference [28] presented an RL based distributed energy management scheme to maximise the profit through energy management and load scheduling without prior information. Another distributed operation strategy was proposed in [29] to operate a community battery energy storage system based on a double deep Q-learning method. In [30], the authors presented a decentralised Markov Decision Process (MDP) to solve an online decentralised and cooperative dispatch problem in order to calculate the approximate Q-value function considering communication delay. These studies, however, did not consider the demand side management of Electric Water Heater (EWH) units. In fact, EWHs are responsible for nearly 30% of the electricity utilised by domestic consumers in winter-dominated climates [31].

The application of RL for demand side management of EWHs started to receive some attention in the literature. AI-Jabery et al. in [32] proposed a fuzzy Q learning to control an EWH, and showed that the proposed algorithm could achieve global convergence. In [33], the proposed Q learning and action dependent heuristic dynamic programming methods are shown to reduce the cost of domestic EWHs energy consumption by approximately 26% and 21%, respectively. Reference [34] presented a batch RL approach to control a cluster of 100 EWHs to decrease the daily cost within a learning period of 45 days. The study in [35] applied fitted Q-iteration algorithm to an EWH to control the heater's ON/OFF actions. It is shown that energy consumption was reduced by 15% in comparison to that when a thermostat controller was used. Somer et al. in [36] proposed a model-based RL approach to optimize the heating cycles of an EWH to maximise the self-consumption of the local PV generation. Six residential buildings were tested and the self-consumption of PV generation was increased substantially. Another RL scheme to optimize the hot water production was presented in [37]. A set of 32 houses in the Netherlands was used, and the energy consumption was reduced by roughly 20% without affecting customers' comfort.

In the above studies, EWHs are considered as a standalone system with their own constraints and they are not considered as an integral part of a larger power network that also includes intermittent RES, limited capacity energy storage systems,

diesel generators, and Information and Communication Technology (ICT) systems. Taking a comprehensive approach that also includes the influence of time-delay is an important aspect to realise reliable smart grids in practice. Moreover, in many islands, the energy tariff is subsidised and fixed, and thus there is no incentive for consumers to change their consumption behaviours. Furthermore, lacking knowledge of each consumer's demand profile makes the centralised control of EWH demand less efficient in reducing total power demand while satisfying individual consumer's comfort requirements. Therefore, this paper proposes an intelligent hierarchically distributed strategy based on RL to control EWH units in isolated microgrids. The distributed controllers use a virtual dynamic tariff that is generated centrally. The virtual tariff can be determined and broadcasted hourly using direct measurement of Battery's SOC. Realising that this method makes the system prone to communication delays and packet loss, an alternative approach that is based on prediction is proposed. This method requires the tariff to be broadcasted only once a day. The main highlights and contributions of the paper can be summarised as follows:

1) This paper proposes a distributed control framework to isolated networks whose energy tariff is fixed (as the case in Ushant Island). The proposed framework uses distributed controllers to optimize 150 electric water heaters independently. The main consideration is that the water consumption habits of each household are different so that the distributed approach can more accurately control the water temperature and reduce diesel energy consumption.
2) The distributed RL controller based on a distributed Q-learning algorithm is adopted. It can learn how to choose actions based on experience and be directly applied in real-time to reduce diesel consumption effectively with different EWH demand profiles. Seven different scenarios of different combinations of RES and energy storage are considered. Simulation results show that the energy consumption of the diesel with RL algorithm is reduced by an average of 8.91%(6.675kW) compared to controlling the EWHs by traditional hysteresis control, the proposed algorithm can support the service provider in optimizing the overall energy operation.
3) A dynamic virtual tariff as a cost indicator is proposed to provide a directive/incentive signal for the local RL based controllers to optimize diesel consumption. To investigate the impact of potential time delays on RL algorithm. Two approaches, a direct measurement (DM) strategy and prediction strategy, are proposed for generating the virtual tariff. The simulation results show that the communication time-delay will produce certain fluctuations during the iterations, and the final convergence results will also be affected. It demonstrates that the prediction strategy allows the framework to execute the algorithm on the basis of ensuring communication quality. Results show that the errors caused by the prediction strategy are negligible.

The rest of this paper is organised as follows. The microgrid network is described in Section II along with the proposed virtual tariff. The electric water heater mathematical model, and the time-delay model are introduced in Section III. The proposed algorithms are introduced in Section IV. In Section V, simulation results for different scenarios are presented. Finally, section VI presents the conclusion.

II. MICROGRID DESCRIPTION AND DYNAMIC VIRTUAL TARIFF

The standalone microgrid under study is shown in Fig. 1. It consists of RES, a battery energy storage system (BESS), a diesel generator, and domestic loads. When the diesel generator is not operating, the battery unit acts as the grid forming unit controlling the bus voltage and frequency, and hence absorbing surplus power and supplying shortage power. However, the battery has a finite capacity and thus when it is fully charged, renewable energy production has to be curtailed. Similarly, when it is fully discharged, either some of the loads have to be shed or the diesel generator has to be dispatched. The required capacity of a BESS is normally determined by a set of various factors, such as uncovered energy demand, and excess renewable energy generation, in addition to the technical and financial constraints. If the battery is to be sized to completely eliminate the need for the diesel generator, i.e., rely 100% on RES, the battery capacity has to be large enough to cover any shortage in energy even if it happens very rarely. This may result in a high capital cost of the battery and, therefore, it is more economical to size the battery to cover 80% or 90% of the renewable energy generation and rely on the diesel generator to cover the rest. On the other hand, DR can play an important role in reducing the diesel usage and the battery size.

A. VIRTUAL TARIFF

The general purpose of any DR is to shift the load demand to time periods when the electricity price is low. However, in many islands like Ushant, the energy tariff is fixed and thus traditional demand response becomes difficult. To deal with this hurdle, a dynamic virtual tariff is proposed to optimize the distributed operation of EWHs independently. This tariff is generated at the energy management system (EMS) based on the surplus/shortage of renewable energy generation and hence the battery SOC and the consumption of diesel.

When the SOC is at its maximum limit, there is surplus in renewable energy. When the SOC is between its maximum and minimum limits, the renewable energy and battery are able to supply power demand. However, when the SOC reaches its minimum limit, there is shortage in energy the diesel generator must be started to cover the deficiency. Therefore, the tariff can be simply divided into three levels to reflect surplus/shortage of renewable energy. The value of proposed tariff has a scale of 1 to 3. When the SOC is at its maximum limit, the tariff is set to level 1. For

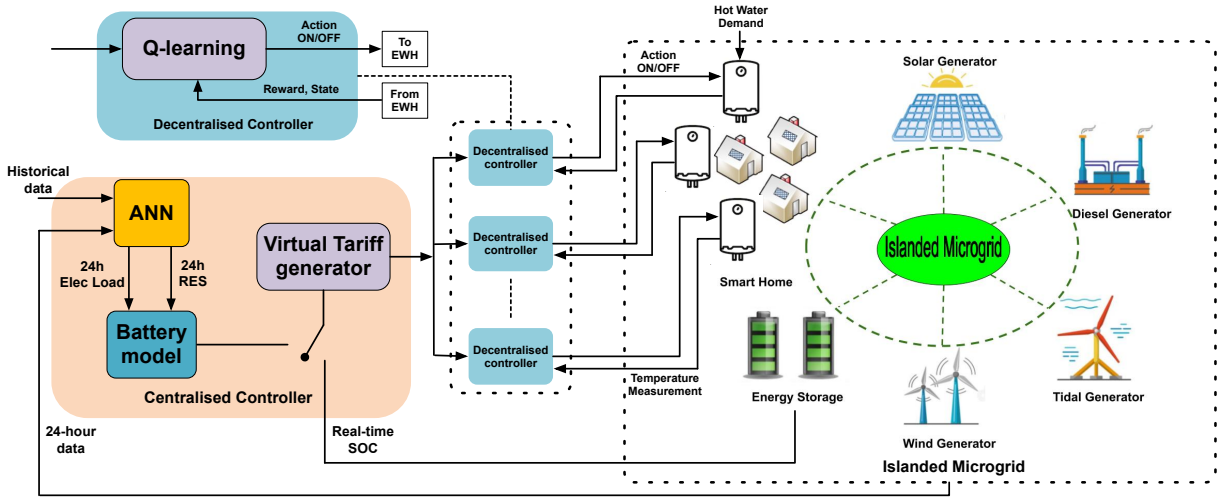

FIGURE 1. Islanded Microgrid and control strategy framework

SOC range from 30% to 100%, the tariff is set to level 2. And when the SOC reaches its minimum value of 30%, the tariff is set to level 3 which means that the battery is fully exhausted, discharge is not allowed, and the diesel generator is operating.

The proposed control structure is shown in Fig. 1. It consists of a centralised controller that generates the virtual tariff at the central EMS and distributed controllers for EWHs. Two strategies for generating the virtual tariff are proposed:

Direct Measurement Strategy: Every hour, the battery SOC is measured directly from the Battery Management System (BMS)and the virtual tariff is then determined, as explained above, and broadcasted to the EWHs' controllers in real time. This method is based on real data but it is prone to communication delays and packet loss.

Prediction Strategy: At the start of each day, the historical data is used to predict the generation of RES and load demand for a 24-hour horizon. Generation of renewable energy sources and load demand can be predicted with high accuracy [38], [39], and thus they are assumed to be known during the optimization process. Two years' historical data is used to train an Artificial Neural Network (ANN) model to predict renewable energy generation and load demand, and is updated every 24 hours. A battery model is then used to calculate the SOC profile for 24 hours. The virtual tariff is then calculated and broadcasted to the distributed controllers. This strategy broadcasts the tariff once a day which will reduce the potential impact of communication delays or packet loss in advance.

Once the tariff is broadcasted, the distributed RL controllers will select appropriate actions to operate the EWHs locally to minimise virtual cost in real-time which will result in a reduction in diesel consumption in the island but at the same time satisfy consumers' requirements in terms of maintaining comfortable water temperature.

III. SYSTEM MODELLING

A. ELECTRIC WATER HEATER MODEL

The thermal model of the EWH describes the dynamic heat-power exchange while considering the inlet cold water and environmental conditions. The dynamic thermal model can be obtained using the Equivalent Thermal Parameter (ETP) approach [40], [41]. When the EWH is ON between the time t and $t+1$, the temperature at $t+1$ can be obtained as:

$$T_{t+1} = (T_t - \tfrac{\beta+Q}{\alpha})e^{-\alpha\Delta t} + \tfrac{\beta+Q}{\alpha}. \tag{1}$$

where $\alpha = \frac{1}{RC}$, $\beta = \frac{T_o}{RC}$ and $R = \frac{1}{UA}$. Q is proportional to the power rating of EWH.

On the other hand, when the EWH is OFF between time t and $t+1$, Q is zero and the temperature at $t+1$ drops due to the thermal loss and inlet cold water.

$$T_{t+1} = (T_t - \tfrac{\beta}{\alpha})e^{-\alpha\Delta t} + \tfrac{\beta}{\alpha}. \tag{2}$$

Consumed hot water is continuously replaced by cold water through the tank inlet. Therefore, the water temperature can be obtained as

$$T_{t+1} = \frac{(M_f - M_t)T_t + T_o M_t}{M_f}. \tag{3a}$$

$$M_t = \rho v A \Delta t. \tag{3b}$$

Combining equations (1) to (3), the mathematical function that describes the dynamics of the EWH can be expressed as

$$T_{t+1} = (\tfrac{(M_f - M_t)T_t + T_o M_t}{M_f} - \tfrac{\beta + \dagger_t Q}{\alpha})e^{-\alpha\Delta t} + \tfrac{\beta+\dagger_t Q}{\alpha}. \tag{4}$$

$$s.t. \quad \forall\, t \in 1, ..., \mathcal{T}$$

$$\dagger_t = \begin{cases} 1 & if \quad ON \\ 0 & if \quad OFF. \end{cases} \tag{5}$$

B. COMMUNICATION DELAY MODEL

In order to investigate the possible impact of communication delay on the RL algorithm, a mathematical delay model is proposed to calculate the delay probability and incorporate it to the proposed approach. From a measurement point of view, the end-to-end delay across a settled path essentially consists of two parts: a deterministic delay D_d and a stochastic delay D_s. The probability density function (PDF) of delay can be written as [42]

$$\begin{aligned}\varphi(t) &= p\varphi_1(t) + q\varphi_1(t) * \varphi_2(t) \\ &= \frac{p}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} + \frac{q\lambda}{\sigma\sqrt{2\pi}} e^{-\lambda t} \int_0^t e^{\lambda u - \frac{(u-\mu)^2}{2\sigma^2}} du.\end{aligned} \tag{6}$$

where $p+q=1$ and $\varphi_1(t) * \varphi_2(t) = \int_0^t \varphi_1(u)\varphi_2(t-u)du$. $\varphi_1(t)$ is the deterministic delay density that can be approximated by $\varphi_1(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$. $\varphi_2(t) = \lambda e^{-\lambda t}$ assumes to follow the exponential distribution by one alternating renewal process with the mean length of the closure periods λ^{-1}.

To determine the time-delay probability, (6) can be recalculated to infer the Cumulative Distribution Function (CDF) of time delay such as

$$\begin{aligned}P(t) &= \int_0^t \varphi(u)du \\ &= \frac{1}{2}\{erf(\frac{\mu}{\sqrt{2}\sigma}) + erf(\frac{t-\mu}{\sqrt{2}\sigma})\} \\ &+ \frac{p-1}{2} e^{\eta}\{erf(\frac{\lambda\sigma^2+\mu}{\sqrt{2}\sigma}) + erf(\frac{t-\lambda\sigma^2-\mu}{\sqrt{2}\sigma})\}.\end{aligned} \tag{7}$$

where $\eta = \frac{1}{2}\lambda^2\sigma^2 + \mu\lambda - \lambda t$ and $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}$. The relative parameters is set as $\mu = 5.3ms$, $\sigma = 0.078$, $p = 0.580$ and $\lambda = 1.39$ [42]. According to (7), the probabilities of different time-delay for each broadcast can be added to analyze the performance.

IV. PROPOSED REINFORCEMENT LEARNING FOR EWH CONTROL

Reinforcement learning is an area of machine learning concerned with how to take actions in an unknown environment so as to maximise a cumulative reward. It learns by modifying an optimization policy in real-time through interacting with the environment and using past experience. The dynamic EWH problem is modelled as a discrete finite MDP. In this model, the EWH operation (ON/OFF) depends on the virtual tariff and the water temperature. RL elements including state and action spaces, reward function, learning and exploration rates, and discount factor are described in detail in the following subsections:

1) State Space

The state variables are time of day (ToD_j), virtual tariff ($Tariff_k$) and water temperature ($Temp_l$).

$$S = \left\{ s|s_{j,k,l} = (ToD_j, Tariff_k, Temp_l) \quad \begin{array}{l} j = 1:J \\ k = 1:K \\ l = 1:L \end{array} \right. \tag{8}$$

where ToD is discretised into J = 144(24x6, every 10 minutes), the virtual tariff is divided into $K = 3$ levels in the range of 1 to 3, and the water temperature is divided into $L = 5$ levels between 55°C and 70°C.

2) Action Space

Action Space is the ON/OFF commands for each EWH

$$A = \{a|(ON, OFF)\} \tag{9}$$

3) Reward

$$r_t = -E_t + L_t. \tag{10}$$

where r_t updates the Q table and to encourage the agent to choose the appropriate action. E_t is based on the virtual tariff and L_t facilitates consumer preferences and comfort requirement:

$$E_t = \left\{ \begin{array}{ll} Power * \Delta t * Tariff & , \quad ON \\ 0 & , \quad OFF \end{array} \right. \tag{11}$$

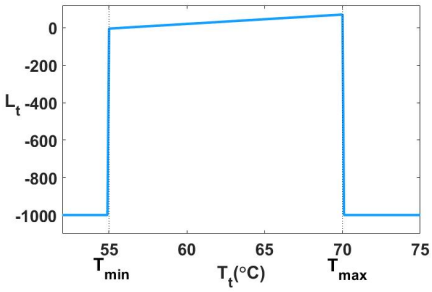

FIGURE 2. Output curve of term L_t for different temperature.

L_t is represented by Fig. 2. It shows high negative penalty for going outside the temperature range of 55 and 70 degrees. It is similar to a coefficient without unit. Furthermore, L_t shows the highest value when the temperature is about 68 degree which reflects consumer preference. Other preferences can be implemented by modifying the reward function. The main purpose of the reward function is used to update the Q table and let the agent to know the quality of different actions. During the iterative process, the reward value will train the RL agent to choose the best action with high probability.

4) Q-learning

In the Q-learning algorithm, an action at a given state is chosen to explore or exploit the future reward value. The Q-value table $Q(s_t, a_t)$ is updated at each iteration. The highest value for each state s in the Q-table corresponds to the highest expected reward after taking action. The optimal updating

policy based on the Bellman equation [43] is expressed as follows:

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha(r(s_t,a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)). \quad (12)$$

where α controls how much previous learning is retained in the update of Q-table. α starts at 0.9, and after 80 days of training it becomes 0.15.

To ensure exploration, an ε-greedy policy is selected [25]. The strategy can either pick an arbitrary action with the probability ε, or take an action corresponding to the maximum value in the Q-value table. ε starts at 0.8 to enable sufficient levels of exploration, and after 80 days of training it becomes 0.01 as the focus moves to exploiting the optimal policy. Note that both α and ϵ decrease with the number of days to ensure sufficient exploration even as the learning process goes on as follows:

$$\alpha = \begin{cases} \alpha_0 & , \ if\ N = 1 \\ \eta\frac{\alpha_0}{\sqrt{N}} & , \ if\ 80 \geq N > 1 \\ 0.15 & , \ if\ N > 80 \end{cases} \quad (13)$$

$$\epsilon = \begin{cases} \epsilon_0 & , \ if\ N = 1 \\ \frac{\epsilon_0}{N} & , \ if\ N > 1 \end{cases} \quad (14)$$

EWH-based Reinforcement Learning Algorithm:
(Prediction and Direct Measurement Strategies)

Initialise all parameters and variables
Select Virtual Tariff generation strategy
%%% **Prediction strategy**
1: **Process** for each day
2: Generate predicted demand and generation via NN
3: Generate the virtual tariff 24-hour ahead
%%% **Direct Measurement strategy**
1: **Process** for each hour
2: Read SOC measurement from the Battery Management System
3: Generate the real-time virtual tariff
%%% **RL for EWH in real-time decision making**
4: **Process** for each agent to do in parallel each hour
5: **Repeat** (for each step in iteration)
6: Choose a_t from current s_t via ε-greedy policy
7: Take action a_t
8: Obtain reward $r(s_t,a_t)$ and next new state s_{t+1}
9: $Q(s_t,a_t) = Q(s_t,a_t) + \alpha(r(s_t,a_t) + \gamma maxQ(s_{t+1},a_{t+1}) - Q(s_t,a_t))$
10: Output the optimal policy
11: **End Process**

The proposed RL Algorithm in the pseudo code shows the detailed DR algorithm, including the prediction strategy and DM strategy. For the RL agent to learn the optimal policy, it has to explore actions that are less rewarding in order to learn from experience. Therefore, it is wise to train the agent offline using historical load/generation data and a mathematical model for the EWH before commissioning. This will avoid operating the real EWHs suboptimally during the learning period.

During offline training, two years' data is provided to train the RL algorithm day by day. Once the training is established, RL can then be commissioned to control EWHs in real time in a model-free fashion; it applies the ON/OFF actions to the real EWH, measures its reward and updates its parameters accordingly. If there is a difference between the model and/or the hot water demand used in the model and those in practice, RL can also adapt to this change thanks to it learning capability.

At the end of each day, the microgrid load/generation data of that day are fed back to the ANN to keep updating the historical data that is used for prediction as shown in Fig. 1.

V. SIMULATION RESULTS

Numerical simulation has been carried out to assess the performance of the proposed DR. The microgrid shown in Fig.1 has been used in this simulation. The training data is obtained from Ushant island in France for the time period of January 1st, 2014 to December 30th, 2015. Another data set from the year of 2016 is used for real-time testing. The load demands of 150 EWH units follow a Poisson distribution, which is proportional to the hourly average household hot water usage is adopted from [44]. A 0.2MW/2MWh Lithium-ion battery storage is used. Seven different renewable energy generation scenarios are explored as shown in Table I [45]. Scenarios 1, 2 and 3 consider wind and solar PV generation while scenarios 4, 5 and 6 consider solar PV and tidal generation. Scenario 7 consists of three types of RES. The diesel generator supplies power only if the load demand cannot be met by RES and the battery. The ANN model is trained by using a long short-term memory (LSTM) network with 256 units. Adam, which is a replacement optimization algorithm for stochastic gradient descent for training deep learning models, is selected as an optimizer with a learning rate of 0.01 via Python. The RL models are established and tested in Matlab.

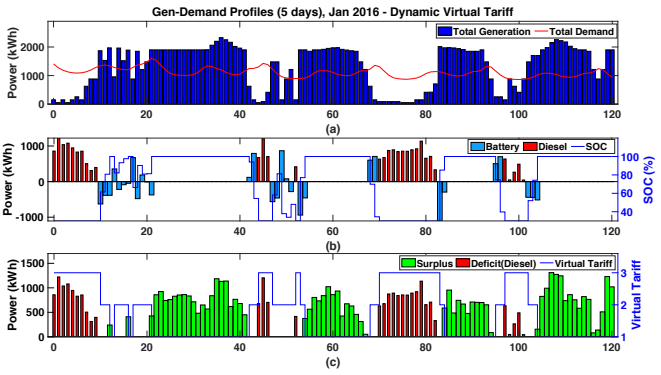

FIGURE 3. Performance of the proposed strategy in Scenario 3.

TABLE 1. Renewable energy resources for seven different scenarios

Scenario Number	Wind	Solar PV	Tidal
1	$300kW$	5 sites, $293.5kW_p$	-
2	$800kW$	20% of rooftops, $3888.93MWh/Annum$	-
3	$2MW$	20% of rooftops, $3888.93MWh/Annum$	-
4	-	5 sites, $293.5kW_p$	Sabella-D10 tidal turbine, $100.4kW$
5	-	20% of rooftops, $3888.93MWh/Annum$	Sabella-D10 tidal turbine, $100.4kW$
6	-	5 sites, $293.5kW_p$	Two Sabella-D10 tidal turbines, $200.8kW$
7	$800kW$	20% of rooftops, $3888.93MWh/Annum$	Sabella-D10 tidal turbine, $100.4kW$

PERFORMANCE OF THE PROPOSED RL-BASED STRATEGY

The generation and demand data for scenario 3 with a 2MWh storage is shown in Fig. 3(a). Battery power and SOC as well as the power from the diesel generator are shown in Fig. 3(b). the virtual tariff is generated by the centralized EMS and is shown in Fig. 3(c) along with surplus and deficit powers. It is clear that the virtual tariff can accurately describe the current state of energy storage and the surplus/deficit of renewable energy, i.e. the state of energy in the whole microgrid.

Off-line Simulation of the RL based Strategy (one day data)

The purpose of this simulation is to demonstrate the ability of RL to achieve optimal performance. According to the one day's virtual tariff, the RL algorithm will update the Q-table and repeat the iterations using the same daily data until convergence is achieved. The energy consumption of an EWH of both strategies is shown in Fig. 4, along with the results obtained using a GA optimization algorithm and the traditional hysteresis control. Two other global optimization approaches, Simulated Annealing (SA) algorithm and Particle Swam (PS) algorithm, are also utilized to verify the experimental results of RL as shown in Table 2. Optimal solution can only be achieved if continuous space/action space is used. Furthermore, in terms of the large search space, the computational cost is expensive and it will also be time-consuming if all state-action pairs need to be visited. The proposed RL can quickly search for sub-optimal solutions and perform real-time control. The results demonstrate that the proposed RL algorithm can reach the optimal results very fast within a few iterations. The energy consumption using the DM strategy and the prediction strategy are 62.53 kWh and 63.73 kWh, respectively. It is very close to the GA result of 61.33 kWh. The energy consumption when the EWH is controlled by the traditional hysteresis controller is 90.67 kWh. However, when the time-delay model is considered, it is shown that the time-delay can lead to large fluctuations and poor convergence. The GA optimizer finds the optimal solution from the simulations and this will always happen unless the GA uses a different model, e.g. a linearised model, or a model without noise. The RL controller converges quickly to the optimal solution, whilst directly interacting with the environment, i.e. without relying on the simulation model.The oscillations are caused by the controller trying to explore new action-state pairs. The superiority of the proposed RL strategy considering the prediction strategy is clearly demonstrated.

TABLE 2. Optimal Results for Different Methods

$Methods$	$Mean$	$Variance$
RL	62.88	3.94e-4
GA	61.38	3.66e-4
SA	61.34	6.09e-4
PS	61.32	1.94e-4

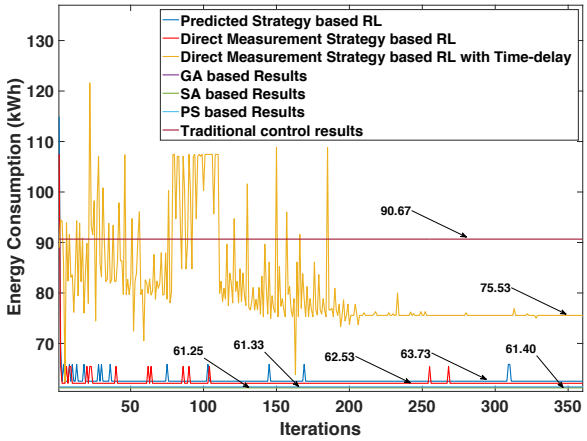

FIGURE 4. Energy consumption of the porposed RL-based strategy compared to the GA and and conventional control strategy.

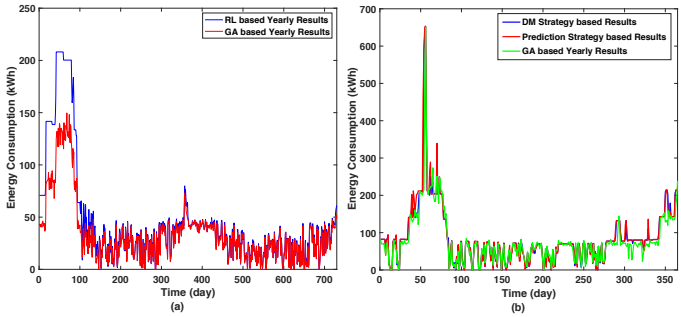

FIGURE 5. (a) Daily energy consumption of training results during two-year period based on the DM strategy, (b) Daily energy consumption of testing results during one-year period (scenario 3).

Off line training of RL (two year data)

Two years of historical generation/demand data from Ushant Island is used to generate the virtual tariff for two years. The virtual tariff according to the direct measurement of SOC

TABLE 3. Impact of Hyper-Parameter

Learning Rate	Results	Epsilon	Results
0.9	128.349	0.005	71.05
0.7	131.64	0.001	70.68
0.5	125.058	0.01	62.53
0.2	65.82	0.05	63.4
0.1	98.73	0.1	80.53

is then used to train the RL agent offline using the EWH mathematical model. Energy consumption is shown in Fig. 5(a). The ability of RL approach to track the optimum cost achieved by the GA algorithm is clear.

Real time control of EWH

The trained RL is used to control 150 EWH units in real time as explained in subsection IV-4. The yearly island load data of 2016 and the resources from scenario 3 are used. The virtual tariff is generated and broadcasted to EWHs in two ways as was explained in section II-A: daily broadcast using ANN prediction of SOC, and hourly broadcast using direct measurement of SOC. The trained RL agent issues the ON/OFF actions on an hourly basis. At the end of each hour, the reward is calculated, the and the next action is chosen. Fig. 5(b) shows the energy consumption for one year along with the results obtained using the GA. Each day has its own optimal consumption value and the proposed RL strategy is able to track this effectively. Both strategies for virtual tariff generation are able to save energy consumption significantly and the results are close to the global optimization policy when time-delay is not considered. Using the RL with DM strategy for generating the virtual tariff can reduce the use of diesel consumption by 8.91% (6.675kW) compared to controlling the EWHs by traditional hysteresis control. If the virtual tariff is generated by the prediction strategy, diesel consumption is reduced by 8.85%, only 0.06% increase compared to DM strategy. This difference, caused by the prediction error, is quite minimal. The advantages for using the ANN are the avoidance of hourly communication with EWH units, and providing customers with the virtual tariff profile in advance.

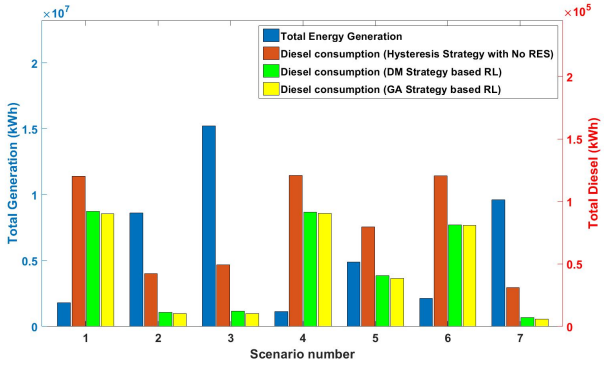

FIGURE 6. Annual energy consumption and generation for seven scenarios in 2016.

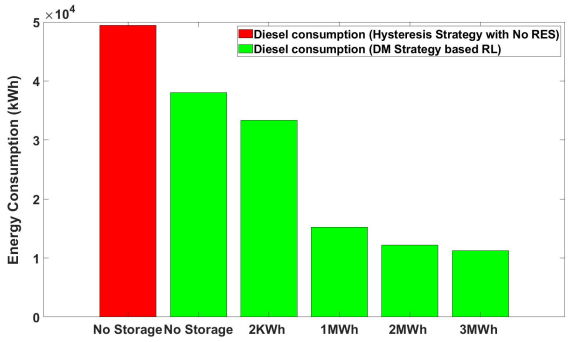

FIGURE 7. Diesel consumption cost for different storage size in scenario 3.

The proposed DR scheme is applied to the seven RES scenarios shown in Table I. The total generation and diesel consumption are presented in Fig. 6 with 150 EWH units being controlled by hysteresis, GA, and direct measurement strategy based RL controllers. It can be noticed that both the RL algorithm and GA algorithm can save diesel cost significantly compared with the traditional hysteresis control, especially in scenario 7. However, the GA requires pre-knowledge of all information in advance and spends a lot of computing resources to get the optimal results. The annual energy consumption when using the RL strategy is very close to that of GA-based strategy in all scenarios. However, the proposed RL algorithm can achieve near optimal results in real-time control with no previous knowledge of EWH models. Furthermore, the yearly summary of the seven different scenarios indicates that RL strategy can cover enough energy demands in scenario 7 and reduce diesel generator consumption significantly. Compared to the other six scenarios, scenario 3 generates up to more than 150 MWh renewable energy generation. However, the total diesel cost in case of hysteresis controlled EWHs shows that there is a substantial surplus of renewable energy not being utilised due to the limitation of the battery size. The results in Fig. 7 for scenario 3 show the diesel consumption cost considering different sizes of batteries. The larger the battery capacity, the more diesel energy is saved. However, considering the battery cost and service life, and the energy consumption of the entire island, the 2WMh capacity energy storage is chosen.

Fig. 8 shows the energy consumption performance of an EWH based on a typical virtual tariff profile (a) when it is controlled by the proposed RL using DM strategy (b) and a simple hysteresis thermostat (c). The virtual price in Fig. 8(a) represents three different prices under three different states (renewable energy only, renewable energy and storage energy only, and diesel consumption only) according to the different electricity prices of different utility companies. It can be seen that the temperatures in both strategies are controlled within the required temperature range ($55°C$ and $70°C$). However, the Fig. 8(b) shows that the RL based strategy can shift the ON commands to periods when the tariff is low. It means

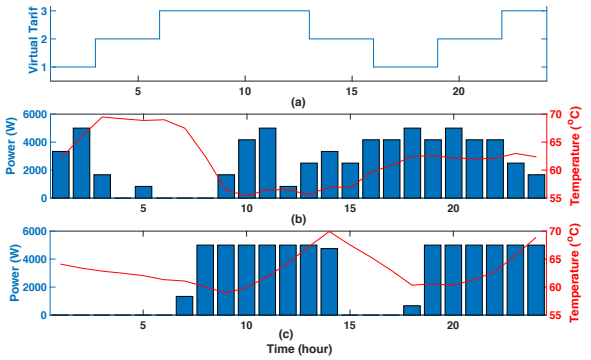

FIGURE 8. Performance comparison of the EWH controller: (a) Example of Virtual Tariff, (b) Proposed RL-based EWH controller, (c) Traditional EWH controller.

that RL agent tends to store energy in the water when there is surplus in energy by keeping the temperature near its maximum. Meanwhile, it can also keep the water temperature just above the minimum during shortage of energy. Furthermore, RL resulted in less total energy consumption compared to that of the hysteresis control approach.

In summary, all the results verify the performance of the proposed DR strategy based on the RL algorithms. It is capable of learning a cost-effective way for EWH management under different conditions, without requiring information about the model in advance.

VI. CONCLUSION

An intelligent distributed real-time DR based on RL has been proposed to manage the demand of 150 EHWs in isolated islands. To overcome the problem of fixed electricity price, an adaptive virtual tariff that reflects the status of the battery and the diesel generator has been generated and used in the reward function of the RL algorithm.Two methods for generating the virtual tariff have been proposed: DM strategy and prediction strategy. Simulation results shows that the prediction strategy is suitable to achieve good performance compared to the DM strategy and it makes the algorithm less dependent on communication time-delay. The prediction strategy can also be used to encourage customers to arrange the use of other electrical equipment in advance to reduce total energy consumption.The performance of the proposed distributed controllers is assessed by simulation which shows the ability of RL to learn the optimal control policy. It is shown that employing the proposed RL algorithm results in an average 8.91%(6.675kW) reduction in the usage of diesel generators for each electric water heater.

However, Q-learning can provide near-optimal solutions effectively which is not friendly to an accurate system model. In future, we seek to consider state-of-the-art reinforcement learning algorithms, such as deep reinforcement learning algorithm and Bayesian reinforcement learning algorithm, to achieve better performance. In addition, the hyper-parameters adjustment has a significant impact on the performance of the RL algorithm and hence we plan to further optimize the parameters by using the state-of-the-art algorithms [46].

REFERENCES

[1] K. Kaygusuz, "Energy for sustainable development: A case of developing countries," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 2, pp. 1116 – 1126, 2012.

[2] U. N. E. Programme, "2020 global status report for buildings and construction towards a zero-emission," *Efficient and Resilient Buildings and Construction Sector. Nairobi.*, 2020.

[3] S. Asumadu-Sarkodie and P. A. Owusu, "Carbon dioxide emissions, gdp, energy use, and population growth: a multivariate and causality analysis for ghana, 1971–2013," *Environmental Science and Pollution Research*, vol. 23, no. 13, pp. 13 508–13 520, Jul 2016.

[4] S. Ahmad, M. Naeem, and A. Ahmad, "Low complexity approach for energy management in residential buildings," *International Transactions on Electrical Energy Systems*, vol. 29, no. 1, p. e2680, 2019, e2680 etep.2680. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/etep.2680

[5] R. Yaqub, S. Ahmad, A. Ahmad, and M. Amin, "Smart energy-consumption management system considering consumers' spending goals (sems-ccsg)," *International Transactions on Electrical Energy Systems*, vol. 26, no. 7, pp. 1570–1584, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/etep.2167

[6] S. Ahmad, M. Naeem, and A. Ahmad, "Unified optimization model for energy management in sustainable smart power systems," *International Transactions on Electrical Energy Systems*, vol. 30, no. 4, p. e12144, 2020, e12144 ITEES-18-0799.R3. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/2050-7038.12144

[7] M. Negnevitsky and K. Wong, "Demand-side management evaluation tool," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 212–222, 2015.

[8] C. Bordin, H. O. Anuta, A. Crossland, I. L. Gutierrez, C. J. Dent, and D. Vigo, "A linear programming approach for battery degradation analysis and optimization in offgrid power systems with solar energy integration," *Renewable Energy*, vol. 101, pp. 417 – 430, 2017.

[9] M. Szypowski, T. Siewierski, and A. Wędzik, "Optimization of energy-supply structure in residential premises using mixed-integer linear programming," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1368–1378, Feb 2019.

[10] L. Li, H. Mu, N. Li, and M. Li, "Economic and environmental optimization for distributed energy resource systems coupled with district energy networks," *Energy*, vol. 109, pp. 947 – 960, 2016.

[11] S. M. Ezzati, G. R. Yousefi, M. M. Pedram, and M. Baghdadi, "Security-constrained unit commitment based on hybrid benders decomposition and mixed integer non-linear programming," in *2010 IEEE International Energy Conference*, Dec 2010, pp. 233–237.

[12] A. Askarzadeh, "A memory-based genetic algorithm for optimization of power generation in a microgrid," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 3, pp. 1081–1089, July 2018.

[13] S. Ahmad, A. Ahmad, M. Naeem, W. Ejaz, and H. S. Kim, "A compendium of performance metrics, pricing schemes, optimization objectives, and solution methodologies of demand side management for the smart grid," *Energies*, vol. 11, no. 10, 2018.

[14] M. Beccali, M. Bonomolo, G. Ciulla, and V. L. Brano, "Assessment of indoor illuminance and study on best photosensors' position for design and commissioning of daylight linked control systems. a new method based on artificial neural networks," *Energy*, vol. 154, pp. 466 – 476, 2018.

[15] G. Ciulla and A. D'Amico, "Building energy performance forecasting: A multiple linear regression approach," *Applied Energy*, vol. 253, p. 113500, 2019.

[16] T. Papadimitriou, P. Gogas, and E. Stathakis, "Forecasting energy markets using support vector machines," *Energy Economics*, vol. 44, pp. 135 – 142, 2014.

[17] P. Moutis, S. Skarvelis-Kazakos, and M. Brucoli, "Decision tree aided planning and energy balancing of planned community microgrids," *Applied Energy*, vol. 161, pp. 197 – 205, 2016.

[18] A. Mechaqrane and M. Zouak, "A comparison of linear and neural network arx models applied to a prediction of the indoor temperature of a building," *Neural Computing & Applications*, vol. 13, no. 1, pp. 32–37, Apr 2004.

[19] B. Kim, Y. Zhang, M. van der Schaar, and J. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2187–2198, 2016.

[20] M. Kamruzzaman, J. Duan, D. Shi, and M. Benidris, "A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources," *IEEE Transactions on Power Systems*, pp. 1–1, 2021.

[21] J. Vlachogiannis and N. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Transactions on Power Systems*, vol. 19, no. 3, pp. 1317–1325, 2004.

[22] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1742–1751, 2012.

[23] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, 2018.

[24] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 1, pp. 1–10, 2019.

[25] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *MIT Press, Cambridge, MA*, 2018.

[26] B. Jiang and Y. Fei, "Smart home in smart microgrid: A cost-effective energy ecosystem with intelligent hierarchical agents," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 3–13, 2015.

[27] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, "Reinforcement learning-based microgrid energy trading with a reduced power plant schedule," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 728–10 737, 2019.

[28] E. Foruzan, L. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5749–5758, 2018.

[29] V. Bui, A. Hussain, and H. Kim, "Double deep q -learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457–469, 2020.

[30] Y. Lan, X. Guan, and J. Wu, "Online decentralized and cooperative dispatch for multi-microgrids," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 1, pp. 450–462, 2020.

[31] A. Moreau, "Control strategy for domestic water heaters during peak periods and its impact on the demand for electricity," *Energy Procedia*, vol. 12, pp. 1074 – 1082, 2011, the Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1876610211019667

[32] K. Al-jabery, D. C. Wunsch, J. Xiong, and Y. Shi, "A novel grid load management technique using electric water heaters and q-learning," in *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Nov 2014, pp. 776–781.

[33] K. Al-jabery, Z. Xu, W. Yu, D. C. Wunsch, J. Xiong, and Y. Shi, "Demand-side management of domestic electric water heaters using approximate dynamic programming," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 5, pp. 775–788, May 2017.

[34] 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 *2014 Power Systems Computation Conference*, Aug 2014, pp. 1–7.

[35] F. Ruelens, B. J. Claessens, S. Quaiyum, B. De Schutter, R. Babuška, and R. Belmans, "Reinforcement learning applied to an electric water heater: From theory to practice," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3792–3800, July 2018.

[36] O. De Somer, A. Soares, K. Vanthournout, F. Spiessens, T. Kuijpers, and K. Vossen, "Using reinforcement learning for demand response of domestic hot water buffers: A real-life demonstration," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Sep. 2017, pp. 1–7.

[37] H. Kazmi, F. Mehmood, S. Lodeweyckx, and J. Driesen, "Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems," *Energy*, vol. 144, pp. 159 – 168, 2018.

[38] A. Motamedi, H. Zareipour, and W. D. Rosehart, "Electricity price and demand forecasting in smart grids," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 664–674, 2012.

[39] M. Tan, S. Yuan, S. Li, Y. Su, H. Li, and F. He, "Ultra-short-term industrial power demand forecasting using lstm based hybrid ensemble learning," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2937–2948, 2020.

[40] S. Katipamula and N. Lu, "Evaluation of residential hvac control strategies for demand response programs (symposium papers - ch06-7 demand response strategies for building systems)," *ASHRAE Transactions*, vol. 112, pp. 535–546, 02 2006.

[41] N. Lu, D. P. Chassin, and S. E. Widergren, "Modeling uncertainties in aggregated thermostatically controlled loads using a state queueing model," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 725–733, May 2005.

[42] J. Zhang, S. Nabavi, A. Chakrabortty, and Y. Xin, "Admm optimization strategies for wide-area oscillation monitoring in power systems under asynchronous communication delays," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 2123–2133, July 2016.

[43] S. Peng, "A generalized dynamic programming principle and hamilton-jacobi-bellman equation," *Stochastics and Stochastic Reports*, vol. 38, no. 2, pp. 119–134, 1992.

[44] L. Gelazanskas and K. A. A. Gamage, "Forecasting hot water consumption in residential houses," *Energies*, vol. 8, no. 11, pp. 12 702–12 717, Nov 2015.

[45] G. S. Matthew, O. W. Fitch-Roy, P. M. Connor, B. Woodman, P. Thies, E. Hussain, H. Mahmood, M. Abusara, X. Yan, and J. Hardwick, "Ice report t2.1.2 - ice general methodology," *INTERREG, University of Exeter*, 2018.

[46] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," 2020.