Contents lists available at ScienceDirect

Journal of Energy Storage

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

Research papers

COST-WINNERS: COST reduction WIth Neural NEtworks-based augmented Random Search for simultaneous thermal and electrical energy storage control

Sven Myrdahl Opalic [a,b,c], Fabrizio Palumbo [d], Morten Goodwin [a], Lei Jiao [a], Henrik Kofoed Nielsen [b], Mohan Lal Kolhe [b,*]

[a] Centre for Artificial Intelligence Research, University of Agder, 4879, Grimstad, Norway
[b] Faculty of Engineering and Science, University of Agder, 4879, Grimstad, Norway
[c] Relog AS, Kongens Gate 16, 7011, Trondheim, Norway
[d] Oslo Metropolitan University, P.O. Box 4 St. Olavs plass, 0130, Oslo, Norway

ARTICLE INFO

Keywords:
Reinforcement learning
COST-WINNERS
Thermal energy storage
Multiple ESS control
Smart warehouse

ABSTRACT

The combination of local renewable energy production, dynamic loads, and multiple energy storage systems with different dynamics requires sophisticated control systems to maximize the energy cost efficiency of the combined energy system. Battery and thermal energy storage systems can be combined to increase the local use of on-site renewable energy, reduce peak power demand, and exploit time-of-use energy pricing. In this paper, we focus on how the augmented random search algorithm and artificial neural networks can be used together to solve an energy cost optimization problem involving the control of a battery energy storage system and a thermal energy storage system at the same time in a smart warehouse. As part of this work, a simulated training environment made using the data from the smart warehouse's operations. In addition to the energy storage systems, the warehouse energy system has integrated a large roof mounted photovoltaic power plant and an industrial-scale cooling system.

The developed solution is able to minimize the energy costs by modulating both energy systems, depending on the situation. Additionally, when it is tested against the state-of-the-art solutions, our developed solution at worst matches performance when the alternative algorithm is allowed to increase training time by a factor of nearly three. On average, our presented solution doubles the performance of the benchmark algorithm with much less computational resource expenditure.

1. Introduction

The current state of global energy supply and demand highlights the need for controllable energy production and storage [1]. There is an increasing demand for robust and responsive electrical and thermal Energy Storage Systems (ESS) [2] as an increasing fraction of the world's energy demand is met by wind and solar power at the expense of fossil-fueled and nuclear power [3]. The building sector represents a natural candidate to deploy an algorithm controlling renewable energy production and storage systems, as buildings are responsible for nearly 40% of global CO_2 emissions [4].

Energy Storage Systems (ESS) can consist of various technologies and be applied in a multitude of ways [5]. From the perspective of the main electrical distribution grid, an important distinction exists between centralized and decentralized ESS. As opposed to decentralized ESS, centralized systems can be directly controlled by the grid operator. However, decentralized ESSs are seen as an important component of a more environmentally friendly energy system, but they come with a new set of challenges [6]. The decentralized systems should monitor the energy market, integrate the control algorithm with market dynamics, and use it to reduce the peak load of the system while also minimizing the costs. In the case of multiple ESSs with different dynamics, such as a combination of a Battery Energy Storage System (BESS) and Thermal Energy Storage (TES), the complexity of the optimization problem further increases.

One approach that is recently gaining a lot of interest in the scientific community as a robust and self-improving method to control building energy systems is Reinforcement Learning (RL) algorithms [7]. RL algorithms can reduce costs by reducing necessary human

[*] Corresponding author.
E-mail addresses: sven.opalic@uia.no (S.M. Opalic), fabrizio@oslomet.no (F. Palumbo), morten.goodwin@uia.no (M. Goodwin), lei.jiao@uia.no (L. Jiao), henrik.kofoed.nielsen@uia.no (H.K. Nielsen), Mohan.L.Kolhe@uia.no (M.L. Kolhe).

https://doi.org/10.1016/j.est.2023.108202
Received 20 December 2022; Received in revised form 2 June 2023; Accepted 26 June 2023
Available online 20 July 2023

Nomenclature

ANN	Artificial Neural Network
ARS	Augmented Random Search
BESS	Battery Energy Storage System
BMS	Building Management System
DC	Direct Current
DDPG	Deep Deterministic Policy Gradient
DPC	Data Predictive Control
DPG	Deep Policy Gradient
DQN	Deep Q-Network
ESS	Energy Storage System
GLPK	GNU Linear Programming Kit
IEMS	Intelligent Energy Management System
MILP	Mixed Integer Linear Programming
MPC	Model Predictive Control
RL	Reinforcement Learning
SAC	Soft Actor-Critic
SOC	State Of Charge
TD3	Twin Delayed Deep Deterministic Policy Gradient
TES	Thermal Energy Storage
TRPO	Trust-Region Policy Optimization

resource expenditure, and risks associated with their behavior can be managed through off-line, data-driven training. Newer RL algorithms often include training Artificial Neural Networks (ANN) to output desired actions or action values, showing improved performance [8–10]. In contrast, Mania et al. [11] showed that the Augmented Random Search (ARS) algorithm could achieve high performance with very little computational resource expenditure by training a simple linear function for action selection with their proposed search algorithm.

In this article we build on the work published in Opalic et al. [12] where we showed that using ANNs for action selection together with the ARS search algorithm improved the agent performance on a BESS control problem. We now propose COST-WINNERS — a novel approach to control, for the first time, both the BESS and TES of a smart warehouse.

Specifically, our contributions in this paper are:

- We implement the ARS [11] RL algorithm, modified with ANNs to encode the agent policy, to simultaneously control TES and BESS energy storage systems.
- We build a data-driven simulated training environment, also modeling the dynamics of the TES.
- Overall, we introduce a novel approach to control both the BESS and TES of a smart warehouse simultaneously to reduce total energy cost. This is important because combining different energy storage systems can lead to improved performance and cost savings but also introduces new challenges due to each system's different dynamics and control requirements.

2. Related work

It was suggested in Xu and Shen [13] an algorithm for optimal control of multiple ESSs using individual custom defined boundaries for energy price. However, the study only features Battery Energy Storage Systems (BESSs) and does not specify how to determine the price boundaries for each system. Zhu et al. [14] examines decentralized ESSs in urban railway applications and suggests multiagent deep Reinforcement Learning (RL) for cooperative control using Q-learning with recurrent ANNs. ANNs are also at the core of Model Predictive Control (MPC) of TES developed by Cox et al. [15]. Zhang et al. [16] propose Soft Actor-Critic (SAC, [17]) to optimize BESS control with multiple energy production facilities. However, the authors have not clarified if the experiment is based on more than a single 24-hour episode and results are only compared with other simpler RL algorithms. Goldsworthy et al. [18] have implemented a cloud-based Model Predictive Control (MPC) battery control algorithm for energy cost reduction at an office building. The system has been operational for a year and achieved an energy cost reduction of 5.5%. Although some of the related work show promising results, we were unable to find any related work that examines advanced control algorithms for energy cost optimization with multiple ESSs with different dynamics, such as the BESS and TES in our smart warehouse.

2.1. Energy optimization in buildings

Similar to the Intelligent Energy Management System (IEMS) implemented in the warehouse and described in our previous work Opalic et al. [12], Sechilariu et al. [19] proposed Mixed Integer Linear Programming (MILP) to optimize energy cost and power flow in a Direct Current (DC) microgrid. Unlike the implemented smart warehouse IEMS, it also features instant power balancing. A hybrid Model, suggested by Huang et al. [20], uses MPC for energy cost optimization in a case-study of an airport terminal. The authors suggest ANNs to account for non-linearity. MPC using hierarchical MPCs to provide thermal comfort and reduce energy cost was suggested in [21]. Smarra et al. [22] propose a data-driven MPC, i.e., Data Predictive Control (DPC), using a random forest algorithm for predictions, claiming that physical models are impractical when considering the unique character and complexity of building-related control systems. On the same line, Rätz et al. [23] also explore data-driven energy system modeling for buildings using RL and MPC. For a twin BESS connected to a wind turbine power plant, Wang et al. [24] suggest MPC. The authors assert greater production dispatchability and increased battery life. To conclude, a review study by Mariano-Hernández et al. [25] determined that the most popular management technique in non-residential buildings is MPC. They come to the conclusion that enabling intelligent control will depend on the building modeling methodology.

2.2. Reinforcement learning

According to Sutton and Barto [26], RL is learning through discovering which actions that increase a reward.

The operational concept of RL is often described as shown in Fig. 1. An agent interacts with an environment, following its internal policy π, by taking actions and receiving feedback from it as reward or penalty. The policy typically gives the agent certain degrees of freedom to choose actions that deviate from the strictest application of the policy. This allows the agent to discover new states and actions that generate higher reward, consequently updating its policy.

Well-known RL algorithms include Q-learning [27] and Deep Q-Networks (DQN) [28]. The Q-learning algorithm maps each state to the expected discounted future value of all the possible actions (Q-value). In Q-learning, the agents' policy is encoded in the Q-table, and the deterministic version of it maximizes Q-value. DQN deploys a deep ANN to compute the Q-values of available discrete actions given an environment state.

The application of RL has been conducted in many different fields, ranging from renewable energy to energy storage, and complex energy systems. An example can be seen in Kuznetsova et al. [29]. The authors developed a simulated microgrid, including a BESS and a wind turbine. The methodology is based on Q-learning taking as inputs the BESS State Of Charge (SOC), energy price, predictions of wind power production, and energy consumption demand. The discrete action space includes three possible BESS actions: charging, discharging, or none. Mbuwir et al. [30] proposed fitted Q-iteration for transfer learning of BESS control to and from systems with comparable properties. Wen et al.

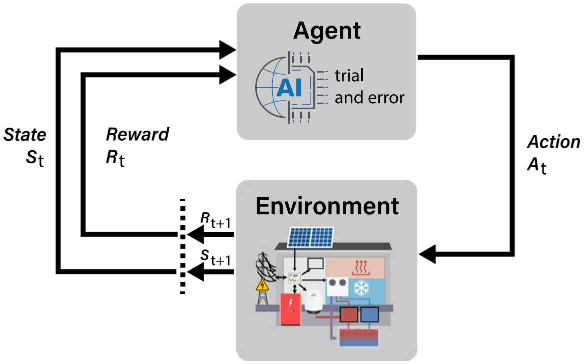


Fig. 1. Interaction between agent and environment.

[31] suggest adopting Q-learning and end-user device utilization for controlling load shifting in modest office and apartment buildings. Additionally, Henze and Schoenmann [32] also used Q-learning for TES control.

Perera and Kamalaruban [7] found that Q-learning is the most common use of RL techniques in the energy research area, even if simpler algorithms are still deployed. Importantly, there are also attempts at exploring state-of-the-art algorithms in the literature. Mocanu et al. [33] propose Deep Policy Gradient (DPG), similar to DQN, for on–off load shifting in the residential sector. Focusing on residential BESS control, a variant of Deep Deterministic Policy Gradient (DDPG) [8] is developed by Wan et al. [34]. Moreover, an improved DQN was implemented also by Cao et al. [9] for BESS arbitrage. This algorithm takes into account a lithium-ion battery degradation model, with discretized action space for full or 50% capacity dynamics together with the stand-by state. Shang et al. [10] combines DQN with bootstrapping and a Monte Carlo tree search for BESS control in a microgrid. However, in all cases except Wan et al. [34], the algorithms work in a discrete domain, having limited action space. In addition, the reward functions are generally complicated and experiment specific. Therefore, most of the approaches mentioned are not ideal for large-scale implementation of IEMS in a multitude of sites using RL.

Brandi et al. [35] explored control of a TES using online deep RL, MPC and offline deep RL. For the online RL controller, energy cost was increased by 160% for a four week period before it converged to comparable behavior to the top performing MPC and offline RL controllers. The study is limited to optimizing electricity cost incurred by the chiller while disregarding overall building energy cost and potential peak power cost.

Wang and Hong [36] conducted a survey of RL application to control technical systems in buildings. The authors argue that established techniques such as MPC requires extensive domain knowledge to properly design and implement, making it less applicable in the building control domain compared with mass production domains such as the automobile industry. Furthermore, Wang and Hong [36] state that RL combined with transfer learning should be further explored for building control.

The authors in Xu et al. [37] propose a combination of RL with differential evolution to reduce energy cost for industrial users with solar power and thermal energy production, as well as BESS and TES, while satisfying local energy demand and trading energy in an energy trading platform.

2.3. Augmented random search

ARS is an optimization of what was named basic random search by Mania et al. [11]. ARS is designed for continuous action space and works with a strictly linear policy matrix, as opposed to other current RL approaches. Moreover, exploration with the ARS is done directly in the parameters of the policy function. In comparison, algorithms such as SAC [17], DDPG [8], TD3 [38], and Trust-Region Policy Optimization (TRPO) [39], also operating in continuous action space, promote action exploration with random noise added to the agents selected action. In the ARS algorithm, random noise is generated and added directly to the policy parameters and tested in the environment. The rewards from N such tests, or rollouts, are then sorted in descending order [11]. The top b directions are used to update the policy according to

$$\theta_{j+1} = \theta_j + \frac{\alpha}{b\sigma_R} \sum_{k=1}^{b} \left[r\left(\pi_{j,(k),+}\right) - r\left(\pi_{j,(k),-}\right) \right] \delta_{(k)}, \tag{1}$$

where θ represents the policy parameters, α represents the learning rate, σ_R is the reward standard deviation, $r(\pi_{j,(k),+})$ and $r(\pi_{j,(k),-})$ are the rewards from rollouts and $\delta_{(k)}$ is the random noise fitted in size to θ. Continuously updated mean and standard deviation of input variables are used to normalize the inputs. Mania et al. [11] managed to achieve outstanding performance while also drastically using less computational resources when tested in a variety of well-known RL benchmark problems.

Table 1
Main components of the smart warehouse energy system.

System	Characteristic value	Unit of measurement
Solar power plant	1000	[kW_p]
BESS	460/200	[kWh/kW]
TES	300/300[a]	[$m^3/kW_{thermal}$]
Cooling plant	1140	[$kW_{thermal}$]
Electric boiler	500	[kW]

[a] At 10 °K temperature difference.

Table 2
Thermal energy storage system characteristics.

Attribute	Values	Unit of measurement
Measurements	L × W × H - 12 × 10 × 2,5	[m]
Volume	300	[m^3]
Average U-value	0.20	[$\frac{W}{m^2K}$]
Ambient temperature	7	[°C]
Storage medium	Water	N/A
Heat exchanger max flow	25	[$\frac{m^3}{h}$]
Heat exchanger temperature loss	2	[°K]

3. Smart warehouse energy system

The energy system in the smart warehouse has previously been described in detail in [12,40,41]. Table 1 lists its main components and a scheme of it is visualized in Fig. 2. In this work we focus mainly on describing the thermal components of the energy system, and specifically the TES.

The main thermal components of the energy system are:

- the cooling plant with cooling energy distribution through evaporators based on direct expansion of carbon dioxide
- heat recovery from the cooling plant with hydronic heating energy distribution
- TES in an insulated firewater tank submerged in the ground.
- cooling for ventilation and server rooms with hydronic cooling energy distribution

The physical characteristics of the tank are listed in Table 2. TES specifications and model parameters are listed in Table 2. Additionally, the energy system also features a BESS, described in detail in [12], that is controlled simultaneously with the TES.

The TES is used to store both heating and cooling energy. Switching between heating and cooling storage, on the other hand, incurs a

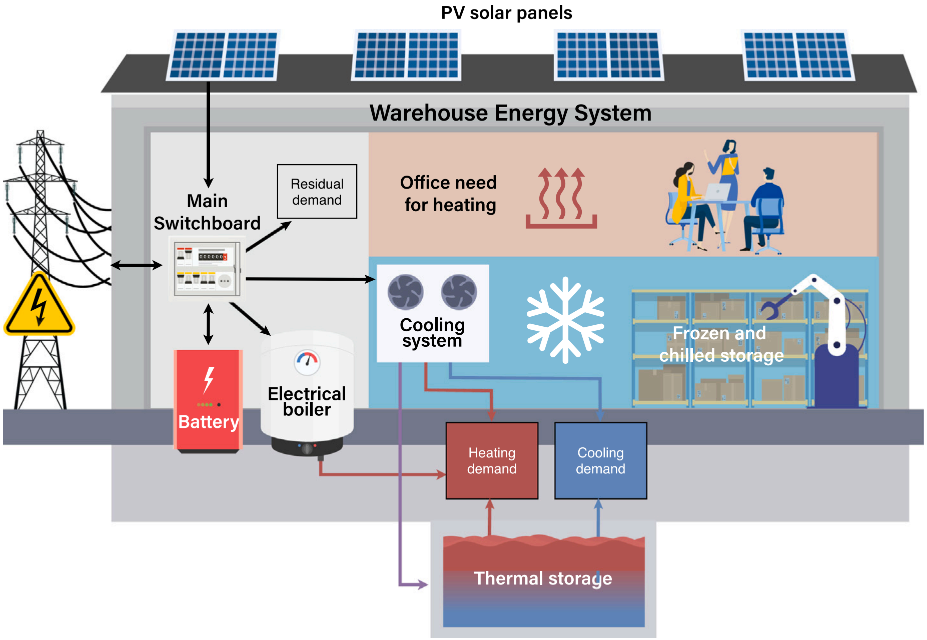


Fig. 2. The smart warehouse energy system with BESS, TES, cooling system and PV power plant. Arrows indicate the direction of energy flow.

significant cost due to the difference in operational temperature levels of the heating and cooling distribution systems at 50 °C and 25 °C, and 9 °C and 15 °C, respectively. Therefore, the TES is used only for heat storage in winter and for cooling storage during summer. For the remainder of this paper, we focus on the TES in heat storage mode. Since the TES is located underground, the ambient temperature also remains relatively stable and is modeled as a constant temperature.

Excess heat is recovered from the cooling plant and can either be directly distributed to cover the warehouse heating demand or stored in the TES, or both. Available excess heat depends on the cooling demand of the refrigerated areas in the building and will vary proportionally to the cooling work done by the cooling plant. If available heat is not sufficient to cover the heating demand, the remaining demand can either be met by discharging stored energy from the TES or by producing heat with an electrical boiler. The boiler can produce heat at an efficiency of around 0.9, whereas using excess heat from the cooling plant only incurs a small cost based on various operating conditions such as internal operating pressure, operational temperature, external cooling demand, and ambient temperature. Recovering and storing excess heat for later discharge can therefore be defined as a time-dependent optimization problem for energy cost reduction.

An IEMS currently controls the on-site ESSs by applying machine learning to predict load and PV solar panel production [42]. Additionally, an optimization algorithm calculates a two-day plan for the BESS and a TES deployment. The local Building Management System (BMS) implements the schedule and updates it hourly. The current IEMS system does not react to live operational data. Every hour the system calculates another two-day schedule, implementing the first hour's actions. Therefore, the system is very dependent on accurate predictions for maximum energy storage and cost reduction. Furthermore, in this scenery, it is challenging to prevent excessive peak power load costs. The magnitude of the challenge is only amplified if we take into account the structure used for the monthly peak power tariff, by the local grid operator: the entire monthly peak power cost is dependent on the highest hourly peak of that month. It is clear then that combining long-term planning with short-term reactions is a key strategy to benefit from the ESS' capability for peak power shaving.

4. Methodology

In this paper, we examine the applicability of the ARS-ANN RL algorithm to a complex energy cost reduction problem through direct control of BESS and TES charging and discharging setpoints in a simulated case-study smart warehouse. Our main research goal is to examine if the ARS-ANN algorithm can efficiently control multiple ESSs with different dynamics and substantially varying degrees of impact on energy cost. The agent is trained in a simulated environment of the smart warehouse, which we mainly designed through the use of data-driven techniques. We have emphasized the use of data-driven techniques as a way to reduce the need for human expertise to design the simulated environment and increase the practical utility of our approach.

4.1. Simulated environment

We have built the simulated environment on operational data using linear and polynomial regression in order to make the simulated environment accessible for result analysis. As this potentially decreases the accuracy of the system model, one could consider building a more accurate model of the environment using deep learning neural networks in an operational scenario. The methodology described in [23] or similar approaches would then be considered. The current version of the simulated environment features an ensemble of models of energy system components and dynamics.

We use a model for the thermal energy storage, production and distribution featuring:

- The heat exchanger temperature loss.
- Temperature loss through heat conduction to surroundings.
- 4 vertical internal temperature levels.

Important components and dynamics of the models for the TES, production, and distribution are the following:

- Operational data of TES charging and discharging compared to setpoint.
- TES storage loss and internal temperature levels.
- Cooling plant electrical power consumption and recoverable excess heat.

A schematic of the TES is included in Fig. 3. The schematic shows the TES in the bottom visualized as a rectangular prism. Physically, the TES is a subterranean concrete basin, insulated on all sides with

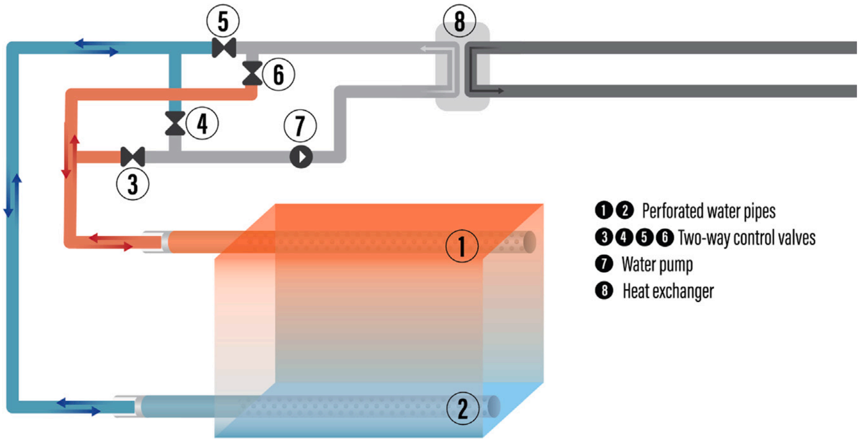
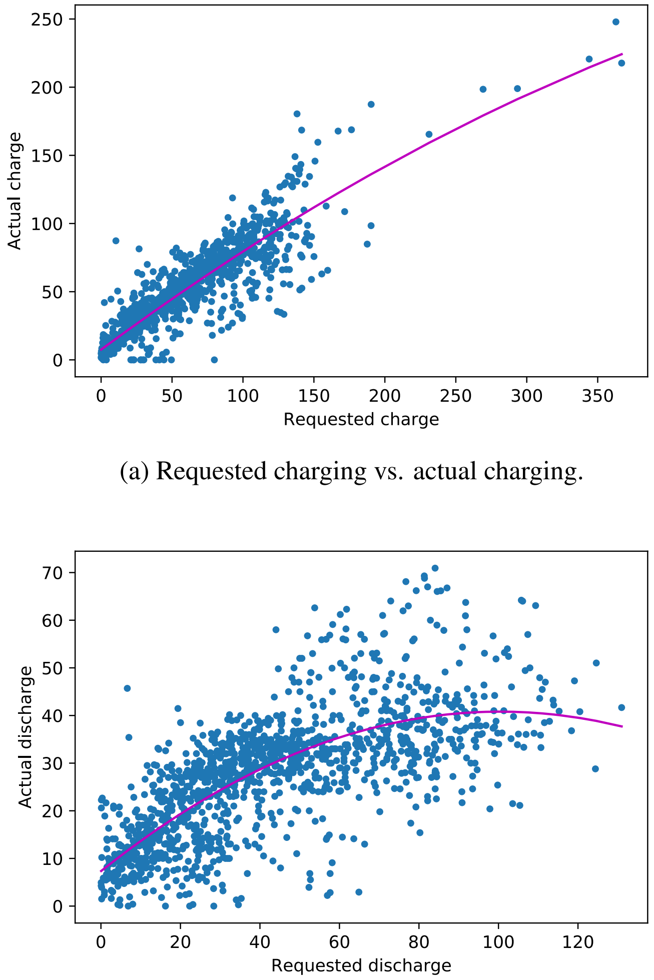


Fig. 3. Thermal energy storage with valves for reversing direction of water flow.

perforated water pipes (1, 2) placed diagonally along opposite walls within. This allows for an even distribution of water flowing into and out of the thermal storage, consistent with a strategy of maintaining water temperature layering inside the tank. The direction of water flowing through the tank can be reversed using an arrangement of four two-way valves (3–6). The TES is physically separated from the main hydronic energy distribution systems by a heat exchanger (8). The flow volume on the TES side of the heat exchanger is automatically balanced with the main hydronic energy distribution system using flow measurements and a frequency controlled pump (7). Our model of the TES includes the ability to reverse the direction of the flow of water such that hotter water is always added to or extracted from the top of the tank and vice versa for colder water. We have not included a model of the heat exchanger due to the physical system automatically balancing volume flow on both sides of the heat exchanger and the observed temperature loss in the heat exchanger is minimal. Modeling the heat exchanger could possibly be considered for future work.

On the secondary side of the heat exchanger, the TES is connected to the hydronic distribution system in two ways (not shown in the figure). Firstly, the TES is connected in parallel with all the thermal heat loads with a modulating two-way control valve that controls the charging according to an external thermal power setpoint. Secondly, the TES can be discharged by circulating the combined return flow through a modulating three-way valve that also responds to an external thermal power setpoint.

However, the dynamics of the hydronic heating system is complicated. We have therefore examined TES operational data in response to charging and discharging set points. The examination shows a high degree of variation between the actual delivered and the requested charge, as well as a non-linear relationship between charging and discharging dynamics. Therefore, we chose to model charging and discharging dynamics with two different functions, using more recent operational data. Charging dynamic is shown in Fig. 4(a), while the discharging dynamic is illustrated in Fig. 4(b). However, we provide a TES action space balanced around the origin of [−100, +100] to the agent interacting with the environment. Actions below 0.2 and above −0.2 are regarded as standby, or no-action. The R^2 score for the charging and discharging functions is 0.83 and 0.53, respectively. A qualitative analysis of the figures highlight a larger spread in the data point for the discharge function. Importantly, although the R^2 for the discharge function is rather low, the goal of this function is to have a simple and explainable model of the TES while discharging. The variation in TES discharging, related to the setpoint, is known to depend on a multitude of other variables when considering a priori and empirical knowledge of the hydronic heating system and is beyond the scope of this paper. A more practical way to model the TES dynamic, with a higher degree of accuracy, is likely through the use of ANN and multiple input variables. However, this would reduce model explainability, and it is not desirable at the current stage.



(a) Requested charging vs. actual charging.

(b) Requested discharging vs. actual discharging.

Fig. 4. Requested TES action vs. actual response with polynomial function fit.

In this article we have implemented the warehouse model described in [40], and configured it to continuously calculate the refrigerant mass flow in the cooling plants. We have fitted a linear regression model, using pressure and mass flow of the refrigerant as inputs and recoverable heat as output. Consequently, this model can be used to find the recoverable heat upper bound at the maximum pressure of 80 bar and at any given refrigerant mass flow.

Moreover, we also model the electric consumption of the cooling plant as a second order polynomial, using refrigerant mass flow and heat recovered as inputs, and the electric consumption as output. The R^2 (R-Squared) score of the electric consumption function is 0.87, while the RMSE is 11.17.

The cooling work, expressed as the refrigerant mass flow, represents the limiting factor for the maximum heat that can be recovered. We model this dynamic with a simple linear function, using as input the refrigerant mass flow and returning as an output the maximum recoverable heat.

Finally, there is a minimum amount of electrical energy required by the cooling plant to keep the storage areas refrigerated. Also in this case we chose a linear model using as input the refrigerant mass flow and returning as an output the least required energy.

The following historical data sources were examined and used as input for the smart warehouse model:

- Total power consumption and local power production.
- Cooling plant power consumption.
- Cooling plant mass flow [40].
- Heating demand.
- TES charging and discharging.
- Energy price for electrical energy bought from and sold to the grid.

4.2. ARS with ANN

In [12], we implement a modified version of the ARS algorithm [11]. We deploy an ANN for policy parametrization in place of the linear function proposed by Mania et al. [11], see Algorithm 1. We thereby modify the processing of inputs to output from a linear to a nonlinear function. More specifically, the ARS algorithm is used to train an ANN to output actions for the TES and BESS with the input being the current state of the environment. We take advantage of the functionality for neural networks already implemented in the RLLIB programming library. Refer to [12] for a detailed explanation of the implemented solution. The algorithm in this article is based on the previously suggested approach.

We use Pyomo [43,44], an open-source Python tool for optimization modeling, with a GNU Linear Programming Kit (GNU Linear Programming Kit (GLPK)) solver to calculate near-optimal solutions for performance comparisons and benchmarking. We feed the GLPK solver with all the information about the training scenario and it attempts to find an optimal solution. However, due to the complex nature of our energy system, we did not attempt to implement the TES in the GLPK solver solution. We examined the operational data and found that the electrical boiler had contributed very little to satisfying the heating demand in the selected time period due to the fact that available excess heat from the cooling system seemed to be sufficient. Reducing energy consumption on the boiler is the main way that the TES can contribute to lower electrical energy consumption during winter operation. We argue that the impact of the TES on the energy cost in the time period we pulled our operational data from is very limited. Adopting the performance of the GLPK solver's control of the BESS as a benchmark is therefore still valid and useful.

5. Scenarios: Results and discussions

In this section, we investigate the application of the ARS-ANN algorithm in a case-study smart warehouse, featuring both electrical (BESS) and thermal (TES) energy storage systems. Therefore, we have the opportunity of analyzing algorithm performance on a complex temporal energy optimization problem. The objective of the algorithm is to reduce energy cost by controlling charging and discharging setpoints of both energy storage systems, BESS and TES.

Algorithm 1 Augmented Random Search with ANN.

```
Set hyperparameters:
    • $\alpha$ - learning rate
    • $n$ - number of directions sampled per iteration
    • $\upsilon$ - exploration noise standard deviation
    • $b$ - number of top-performing directions to use
Run algorithm 2 to initialize policy parameters $\theta_j$, i.e. ANN weights
Initialize:
    • Mean - $\mu_0 = 0 \in \mathbb{R}^{inputs}$
    • Covariance - $\Sigma_0 = \mathbf{I}_n \in \mathbb{R}^{inputsxinputs}$
while ending condition not satisfied do
    Sample $\delta_1, \delta_2, ..., \delta_N$ of the same size as $\theta_j$, with i.i.d. standard normal entries.
    Normalize input values $x$ with $x_{normalized} = diag(\Sigma_j)^{-\frac{1}{2}}(x - \mu_j)$. Collect $2N$ rollouts of horizon $H$ and their corresponding rewards using noise modified ANN policies $\pi_{j,k,+}$ and $\pi_{j,k,-}$, where the $\upsilon\delta_k$ exploration noise is added to the weight parameters $\theta_j$ of the ANN for $\pi_{j,k,+}$ and subtracted from $\theta_j$ for $\pi_{j,k,-}$ with $k \in \{1, 2, ..., N\}$.
    Sort the directions $\delta_k$ by $\max\{r(\pi_{j,k,+}), r(\pi_{j,k,-})\}$, denote by $\delta_{(k)}$ the $k$-th largest direction, and by $\pi_{j,(k),+}$ and $\pi_{j,(k),-}$ the corresponding policies.
    Make the update step for the ANN weights: $\theta_{j+1} = \theta_j + \frac{\alpha}{b\sigma_R} \sum_{k=1}^{b} [r(\pi_{j,k,+}) - r(\pi_{j,k,-})]\delta_k$, where the standard deviation of the $2b$ rewards for the policy update is $\sigma_R$.
    Set the mean and covariance, $\mu_{j+1}, \Sigma_{j+1}$, of the $2NH(j+1)$ training states encountered.
    $j \leftarrow j + 1$.
end while
```

Algorithm 2 ANN for ARS in RLLIB.

```
Set hyperparameters:
    • $\theta^{hl}$ - ANN hidden layers.
    • $\theta^{nu}$ - number of neurons in each hidden layer.
    • $\theta^{af}$ - list of activation function for each layer.
Initialize: $j = 0$, policy parameters $\theta_j$ of shape defined by $\theta^{hl}$ and $\theta^{nu}$ and random values $X$ from $N(\mu_\theta, \sigma_\theta^2)$ normal distribution of mean $\mu_\theta = 0$ and variance $\sigma_\theta^2 = 1$, multiplied by standard deviation $\sigma = 1.0$ for the hidden layers and $\sigma = 0.1$ for the output, divided by the square root of the random value $X^{hl,nu}$, $\theta_j^{hl,nu} = X^{hl,nu}\frac{\sigma}{\sqrt{X}}$.
```

5.1. Scenario I — proof of concept

In scenario I (i.e. first experiment), we apply the ARS-ANN agent to control both BESS and TES for a random 48-hour episode. Our results clearly indicate that the agent is able to find a near-optimal value for BESS charging such that the peak power cost is reduced to a minimum. Change in cooling plant electrical consumption due to control of the TES is shown in Fig. 5, whereas total energy consumption and ESS actions performed by the ARS-ANN agent are shown in Fig. 6. As we can observe in Fig. 6 the maximum hourly energy consumption is flattened, by utilization of the ESS, to around 512 kW, compensating for the consumption peak at almost 600 kW that would occur in the baseline consumption and contributing mainly in the reducing peak power tariff cost.

The agent took advantage of the TES, when heating was required, to reduce the electrical energy required by the cooling plant. It is relevant to mention that the heating demand was very low during the random episode used for experiment one. However, the ARS-ANN agent was

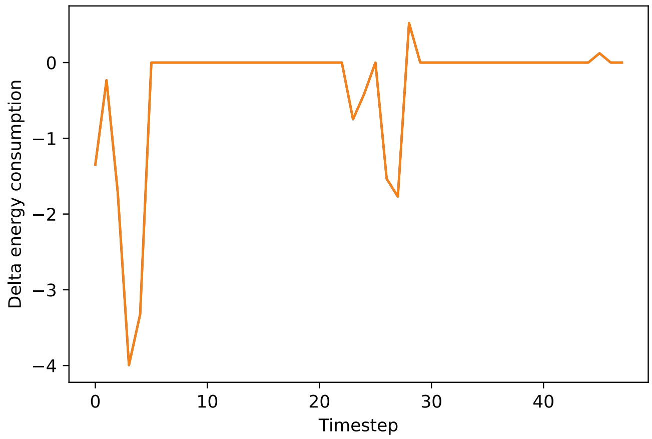


Fig. 5. Change in electrical energy consumption for the cooling system due to ARS-ANN agent TES control.

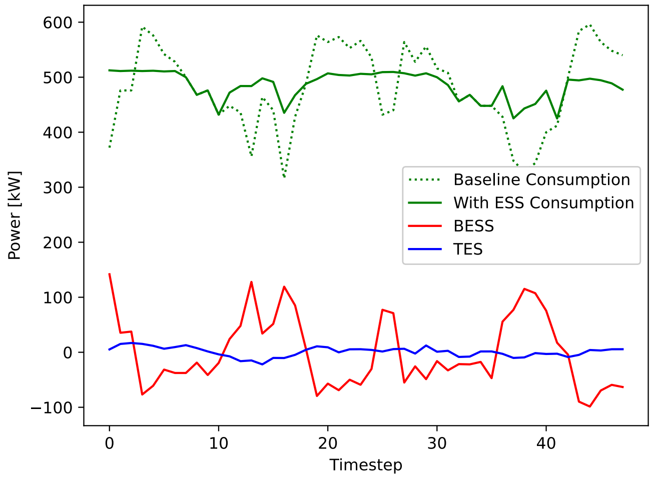


Fig. 6. Total energy consumption and ARS-ANN agent ESS utilization in experiment one.

Table 3
Results for 10 seeded trials for ARS-ANN vs GLPK — battery only.

Trial	GLPK — Battery only	ARS-ANN Result	Percent of GLPK
1	4910	5046	103%
2	7115	7106	100%
3	7540	7498	99%
4	298	361	121%
5	643	639	100%
6	7117	7109	100%
7	5861	5864	100%
8	3771	3780	100%
9	640	641	100%
10	6652	3233	49%

still able to find and store excess heat when there was no cost induced, and then in turn used this to partially reduce electrical consumption by discharging when necessary. Doing this, the agent was able to minimize cooling system energy demand when heat was in demand.

5.2. Scenario II — seeded trials and benchmarking

To better quantify the performance of the ARS-ANN agent, we compare it with a GLPK optimization solver in multiple seeded trials, as well as benchmark it with other state-of-the-art RL algorithms. The GLPK will be controlling solely the BESS, with perfect information, and the comparison will be done for 10 seeded trials. Opposed to the GLPK, the ARS-ANN agent will have control of both BESS and TES. We have decided that comparing performance to an optimization algorithm, with perfect information, of simultaneous BESS and TES control is out of the scope of this paper due to the complexity. Additionally, the operational data used to pull random seeded trials is from early winter where the potential cost reduction of optimal TES control is minor compared with BESS control. There are two main reasons behind this choice of time period. Firstly, this was the time period with the most available data requiring minimal amounts of data cleaning. Secondly, we decided that observing how the algorithm performs in controlling multiple systems with vastly different impact on the result would be of interest.

Table 4
Results for 10 seeded trials with state-of-the-art RL algorithms, compared as a percentage to ARS-ANN results from Table 3.

Trial	SAC		TD3	
	Reward	Percentage ARS-ANN	Reward	Percentage ARS-ANN
1	13	0.3%	346	7%
2	7083	99.7%	290	4%
3	7147	95.3%	43	1%
4	141	39.1%	−62	−17%
5	86	13.4%	76	12%
6	1246	17.5%	305	4%
7	133	2.3%	55	1%
8	3772	99.8%	21	1%
9	728	113.5%	232	36%
10	691	21.4%	−338	−10%

The results of the simulation are displayed in Table 3. We observe that for the majority of the trials, the energy cost reduction of the ARS-ANN with both BESS and TES control either meets or exceeds the cost reduction of the GLPK with BESS control only. For trial 10, the algorithm seems to get stuck in a local optima where it charges the battery too aggressively on the first timestep. Additional research is required to explore why this happens and how it can be avoided in the future. In the 4th seeded trial we observe that the ARS-ANN outperforms GLPK by 21%. In this trial, the potential of cost reduction using the BESS is quite low due to a relatively low baseline peak power cost. Finally, we compare results for the SAC and TD3 RL algorithms to the ARS-ANN algorithm solution, shown in Table 4. In Table 4 the results for SAC and TD3 are compared to the results for ARS-ANN from Table 3. Here, we can observe that TD3 seems to get stuck around original while SAC actually performs reasonably well and even exceeds ARS-ANN in a single trial, finally achieving an average performance of 50% compared with ARS-ANN. However, on a reasonable time frame of running the algorithms for about a week of training time on 6 GPU's and 96 CPU's, both SAC and TD3 achieved similar results. It was only after increasing SAC training time, by a factor of 3, to a total of more than 3 weeks that these results could be achieved. Also, the SAC algorithm results were not stable in the sense that the performance does not stabilize at a high performance. In fact, it drops off entirely in most cases. The results in Table 4 include the maximum award achieved during each training session.

We also ran the seeded trials for the original ARS algorithm to quantify the improvement represented by ARS-ANN. The results showed that ARS performed at an average of 68% compared to GLPK over the 10 trials and hence was outperformed by ARS-ANN by almost 30 percentage points.

5.3. Discussion

In this paper, we examine the applicability of the ARS-ANN RL algorithm to a complex energy cost reduction problem by direct control of BESS and TES charging and discharging setpoints in a simulated environment of an operational smart warehouse.

To evaluate our solution, we use a GLPK optimization solver, controlling only a BESS, as a benchmark. We have decided not to include the TES in the GLPK solver for two main reasons: (i) our initial data analysis demonstrated a marginal impact of the TES, and (ii) its complex thermal dynamics. We argue that for this work a GLPK solver with BESS represents a sufficient approximation to a good solution. We show that for nine out of ten of our seeded trials, the algorithm meets or exceeds the performance of a GLPK optimization solver controlling the BESS only, while given perfect information. For the single trial where it only performs at around 50% of the GLPK, the algorithm seems to get stuck in a local optimum which is to be further explored in future research.

We also compare our solution to state-of-the-art RL algorithms, showing an average of 100% performance increase compared to the SAC algorithm. However, the SAC algorithm was able to match or slightly exceed the performance of ARS-ANN in a few seeded trials when SAC training time was increased by a factor of 3. Further, the best results for SAC were not maintained as the training progressed, meaning that the performance declined after briefly achieving the highest performance for each training session. These "sparks of brilliance" could perhaps be leveraged in some way in future research. It would be of interest, for future work, to investigate possible solutions combining ARS-ANN and SAC for managing BESS and TES.

It is essential to mention that, due to time constraints and a lack of additional data, we only tested our approach in scenarios in which the heating demand was limited. It would be of interest, in future studies, to explore a broader landscape of scenarios, with higher heating demand, to evaluate the general efficacy of the method.

6. Conclusions

We demonstrate that we are able to minimize energy cost in the considered warehouse. We are able to model the dynamics of the TES and to use it in combination with BESS, controlled simultaneously by the ARS-ANN agent.

We demonstrate that by combining BESS and TES with the presented ARS-ANN agent, the agent was able to stabilize maximum energy consumption and thereby reducing the peak power cost. Additionally, the agent was able to exploit the TES when the heat was in demand to reduce the required electrical energy consumption by the cooling plant and electrical boiler.

To conclude, we propose a novel approach to control both the BESS and TES of a smart warehouse simultaneously to reduce total energy cost. This is important because combining different energy storage systems can lead to improved performance and cost savings but also introduces new challenges due to each system's different dynamics and control requirements. The results conclusively show that ARS-ANN outperforms comparable RL algorithms, achieving similar performance to an optimization algorithm controlling the BESS with perfect information.

Declaration of competing interest

All authors declare there is no conflict of interest.

Data availability

The data that has been used is confidential.

Acknowledgment

This work is the result of a collaboration between the Relog AS and University of Agder and it's Centre for Artificial Intelligence Research (CAIR). The work is also partially supported by the Norwegian Research Council.

References

[1] IEA, World energy outlook 2021, 2021, https://www.iea.org/reports/world-energy-outlook-2021.

[2] A.A. Kebede, T. Kalogiannis, J. Van Mierlo, M. Berecibar, A comprehensive review of stationary energy storage devices for large scale renewable energy sources grid integration, Renew. Sustain. Energy Rev. 159 (2022) 112213, http://dx.doi.org/10.1016/j.rser.2022.112213, URL https://www.sciencedirect.com/science/article/pii/S1364032122001368.

[3] J. Buongiorno, J.E. Parsons, D.A. Petti, J. Parsons, The future of nuclear energy in a carbon-constrained world, Mass. Inst. Technol. Energy Initiative (MITEI) (2019).

[4] IEA, Energy efficiency 2020, 2020, https://www.iea.org/reports/energy-efficiency-2020.

[5] O. Palizban, K. Kauhaniemi, Energy storage systems in modern grids—Matrix of technologies and applications, J. Energy Storage 6 (2016) 248–259, http://dx.doi.org/10.1016/j.est.2016.02.001, URL https://www.sciencedirect.com/science/article/pii/S2352152X1630010X.

[6] P.M. Bögel, P. Upham, H. Shahrokni, O. Kordas, What is needed for citizen-centered urban energy transitions: Insights on attitudes towards decentralized energy storage, Energy Policy 149 (2021) 112032.

[7] A. Perera, P. Kamalaruban, Applications of reinforcement learning in energy systems, Renew. Sustain. Energy Rev. 137 (2021) 110618, http://dx.doi.org/10.1016/j.rser.2020.110618, URL https://www.sciencedirect.com/science/article/pii/S1364032120309023.

[8] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, 2015, arXiv:1509.02971.

[9] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, K. Li, Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model, IEEE Trans. Smart Grid 11 (5) (2020) 4513–4521, http://dx.doi.org/10.1109/TSG.2020.2986333.

[10] Y. Shang, W. Wu, J. Guo, Z. Ma, W. Sheng, Z. Lv, C. Fu, Stochastic dispatch of energy storage in microgrids: An augmented reinforcement learning approach, Appl. Energy 261 (2020) 114423, http://dx.doi.org/10.1016/j.apenergy.2019.114423, URL https://www.sciencedirect.com/science/article/pii/S0306261919321105.

[11] H. Mania, A. Guy, B. Recht, Simple random search provides a competitive approach to reinforcement learning, 2018, arXiv:1803.07055.

[12] S.M. Opalic, M. Goodwin, L. Jiao, H.K. Nielsen, M.L. Kolhe, Augmented random search with artificial neural networks for energy cost optimization with battery control, J. Clean. Prod. (2022) 134676, http://dx.doi.org/10.1016/j.jclepro.2022.134676, URL https://www.sciencedirect.com/science/article/pii/S0959652622042482.

[13] Y. Xu, X. Shen, Optimal control based energy management of multiple energy storage systems in a microgrid, IEEE Access 6 (2018) 32925–32934, http://dx.doi.org/10.1109/ACCESS.2018.2845408.

[14] F. Zhu, Z. Yang, F. Lin, Y. Xin, Decentralized cooperative control of multiple energy storage systems in urban railway based on multiagent deep reinforcement learning, IEEE Trans. Power Electron. 35 (9) (2020) 9368–9379, http://dx.doi.org/10.1109/TPEL.2020.2971637.

[15] S.J. Cox, D. Kim, H. Cho, P. Mago, Real time optimal control of district cooling system with thermal energy storage using neural networks, Appl. Energy 238 (2019) 466–480, http://dx.doi.org/10.1016/j.apenergy.2019.01.093, URL https://www.sciencedirect.com/science/article/pii/S0306261919300911.

[16] B. Zhang, W. Hu, D. Cao, T. Li, Z. Zhang, Z. Chen, F. Blaabjerg, Soft actor-critic–based multi-objective optimized energy conversion and management strategy for integrated energy systems with renewable energy, Energy Convers. Manage. 243 (2021) 114381.

[17] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018, CoRR abs/1801.01290 URL http://arxiv.org/abs/1801.01290.

[18] M. Goldsworthy, T. Moore, M. Peristy, M. Grimeland, Cloud-based model-predictive-control of a battery storage system at a commercial site, Appl. Energy 327 (2022) 120038, http://dx.doi.org/10.1016/j.apenergy.2022.120038, URL https://www.sciencedirect.com/science/article/pii/S0306261922012958.

[19] M. Sechilariu, B.C. Wang, F. Locment, Supervision control for optimal energy cost management in DC microgrid: Design and simulation, Int. J. Electr. Power Energy Syst. 58 (2014) 140–149, http://dx.doi.org/10.1016/j.ijepes.2014.01.018, URL https://www.sciencedirect.com/science/article/pii/S0142061514000313.

[20] H. Huang, L. Chen, E. Hu, A new model predictive control scheme for energy and cost savings in commercial buildings: An airport terminal building case study, Build. Environ. 89 (2015) 203–216, http://dx.doi.org/10.1016/j.buildenv.2015.01.037, URL https://www.sciencedirect.com/science/article/pii/S0360132315000530.

[21] V. Lešić, A. Martinčević, M. Vašak, Modular energy cost optimization for buildings with integrated microgrid, Appl. Energy 197 (2017) 14–28, http://dx.doi.org/10.1016/j.apenergy.2017.03.087, URL https://www.sciencedirect.com/science/article/pii/S0306261917303276.

[22] F. Smarra, A. Jain, T. de Rubeis, D. Ambrosini, A. D'Innocenzo, R. Mangharam, Data-driven model predictive control using random forests for building energy optimization and climate control, Appl. Energy 226 (2018) 1252–1272, http://dx.doi.org/10.1016/j.apenergy.2018.02.126, URL https://www.sciencedirect.com/science/article/pii/S0306261918302575.

[23] M. Rätz, A.P. Javadi, M. Baranski, K. Finkbeiner, D. Müller, Automated data-driven modeling of building energy systems via machine learning algorithms, Energy Build. 202 (2019) 109384.

[24] B. Wang, G. Cai, D. Yang, Dispatching of a wind farm incorporated with dual-battery energy storage system using model predictive control, IEEE Access 8 (2020) 144442–144452, http://dx.doi.org/10.1109/ACCESS.2020.3014214.

[25] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, F. Santos García, A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis, J. Build. Eng. 33 (2021) 101692, http://dx.doi.org/10.1016/j.jobe.2020.101692, URL https://www.sciencedirect.com/science/article/pii/S2352710220310627.

[26] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, second ed., The MIT Press, 2018, URL http://incompleteideas.net/book/the-book-2nd.html.

[27] C.J.C.H. Watkins, Learning from Delayed Rewards (Ph.D. thesis, Cambridge University, Cambridge, England), King's College, Cambridge United Kingdom, 1989.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M.A. Riedmiller, Playing atari with deep reinforcement learning, 2013, CoRR abs/1312.5602 URL http://arxiv.org/abs/1312.5602.

[29] E. Kuznetsova, Y.-F. Li, C. Ruiz, E. Zio, G. Ault, K. Bell, Reinforcement learning for microgrid energy management, Energy 59 (2013) 133–146, http://dx.doi.org/10.1016/j.energy.2013.05.060, URL http://www.sciencedirect.com/science/article/pii/S0360544213004817.

[30] B.V. Mbuwir, F. Ruelens, F. Spiessens, G. Deconinck, Battery energy management in a microgrid using batch reinforcement learning, Energies 10 (11) (2017) 1846.

[31] Z. Wen, D. O'Neill, H. Maei, Optimal demand response using device-based reinforcement learning, IEEE Trans. Smart Grid 6 (5) (2015) 2312–2324.

[32] G.P. Henze, J. Schoenmann, Evaluation of reinforcement learning control for thermal energy storage systems, HVAC R Res. 9 (3) (2003) 259–275, http://dx.doi.org/10.1080/10789669.2003.10391069, arXiv:https://www.tandfonline.com/doi/pdf/10.1080/10789669.2003.10391069 URL https://www.tandfonline.com/doi/abs/10.1080/10789669.2003.10391069.

[33] E. Mocanu, D.C. Mocanu, P.H. Nguyen, A. Liotta, M.E. Webber, M. Gibescu, J.G. Slootweg, On-line building energy optimization using deep reinforcement learning, IEEE Trans. Smart Grid 10 (4) (2019) 3698–3708, http://dx.doi.org/10.1109/TSG.2018.2834219.

[34] Z. Wan, H. Li, H. He, Residential energy management with deep reinforcement learning, in: 2018 International Joint Conference on Neural Networks, IJCNN, 2018, pp. 1–7, http://dx.doi.org/10.1109/IJCNN.2018.8489210.

[35] S. Brandi, M. Fiorentini, A. Capozzoli, Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management, Autom. Constr. 135 (2022) 104128.

[36] Z. Wang, T. Hong, Reinforcement learning for building controls: The opportunities and challenges, Appl. Energy 269 (2020) 115036, http://dx.doi.org/10.1016/j.apenergy.2020.115036, URL https://www.sciencedirect.com/science/article/pii/S0306261920305481.

[37] Z. Xu, G. Han, L. Liu, M. Martínez-García, Z. Wang, Multi-energy scheduling of an industrial integrated energy system by reinforcement learning-based differential evolution, IEEE Trans. Green Commun. Netw. 5 (3) (2021) 1077–1090.

[38] S. Fujimoto, H. van Hoof, D. Meger, Addressing function approximation error in actor-critic methods, 2018, CoRR abs/1802.09477 URL http://arxiv.org/abs/1802.09477.

[39] J. Schulman, S. Levine, P. Moritz, M.I. Jordan, P. Abbeel, Trust region policy optimization, 2015, CoRR abs/1502.05477 URL http://arxiv.org/abs/1502.05477.

[40] S.M. Opalic, M. Goodwin, L. Jiao, H.K. Nielsen, Á.Á. Pardiñas, A. Hafner, M.L. Kolhe, ANN modelling of CO2 refrigerant cooling system COP in a smart warehouse, J. Clean. Prod. 260 (2020) 120887, http://dx.doi.org/10.1016/j.jclepro.2020.120887, URL https://www.sciencedirect.com/science/article/pii/S0959652620309343.

[41] S.M. Opalic, M. Goodwin, L. Jiao, H.K. Nielsen, M. Lal Kolhe, A deep reinforcement learning scheme for battery energy management, in: 2020 5th International Conference on Smart and Sustainable Technologies, SpliTech, 2020, pp. 1–6, http://dx.doi.org/10.23919/SpliTech49282.2020.9243797.

[42] G. Marton, et al., MIP in Demand Side Response (Master thesis), Gergely Marton, 2019.

[43] W.E. Hart, J.-P. Watson, D.L. Woodruff, Pyomo: modeling and solving mathematical programs in Python, Math. Program. Comput. 3 (3) (2011) 219–260.

[44] M.L. Bynum, G.A. Hackebeil, W.E. Hart, C.D. Laird, B.L. Nicholson, J.D. Siirola, J.-P. Watson, D.L. Woodruff, Pyomo–Optimization Modeling in Python, Vol. 67, third ed., Springer Science & Business Media, 2021.