POLITECNICO DI TORINO

Master's Degree in Ingegneria Energetica e Nucleare

Master's Thesis

Advanced control strategies for the management of energy storages in multi-energy buildings

Supervisors

Prof. Alfonso CAPOZZOLI

Prof. Silvio BRANDI

Candidate

Antonio GALLO

07/2021

Abstract

Buildings account about a third of the total global final energy use and it is expected to grow further in the next 30 years due to the growth in population, higher indoor comfort level demand, and longer time spent inside buildings. Heating, Ventilation and Air Conditioning (HVAC) systems constitute the major source of energy consumption, thus researchers are trying to develop more efficient control strategies to manage their operation. At the time of this work, the research efforts have been focused on Model Predictive Control (MPC), but the highly diversified building stock and the need of accurate model have slowed down the advancement. New buildings are embracing more advanced enabling technologies for Building Energy Management Systems such as Internet of Things and Cloud Computing, allowing the upstream of information. The so-called Big Data are generated every day and stand for the opportunity to enhance local and supervisory management of the energy systems. Soft-control relies on that data to make predictions, reveal patterns in energy consumption, cluster buildings and for building control, thanks to the computational power given by present-day technologies. Reinforcement Learning (RL) is a promising technique for solving complex non-linear problems, and together with MPC can potentially establish as the state-of-art technologies in the building control field. In this work, Deep Reinforcement Learning is used to manage thermal energy storage in a multi-energy building in Turin during the cooling season. The algorithm is a Soft Actor-Critic. The energy system comprises of a cooling system

with cold-water storage tank, and a Photovoltaic field with Li-ion battery. This work aims at describing how an advanced control strategy affects the design of the thermal and electrical storage. The RL agent schedules the tank operation to minimize the energy cost and it is compared to a benchmark which charges the tank when price is low and discharges it during high price hours. The environment is simulated for several configurations of different Thermal Energy Storage and Battery Energy Storage Systems (BESS) sizes. The results clearly indicate that the RL performs better than the benchmark both in terms of energy consumption and energy cost. Moreover, it guarantees much higher level of self-sufficiency and self-consumption, hence it can be considered that RL is a viable alternative to the sizing up of the energy storage when it comes to nearly Zero Energy Buildings design. As a result, large BESS with unbearable cost can be avoided, even if Variable Renewable Energy Sources are considered. By making use of RL, the grid is less involved in building operation, which is also desirable when seeking for building flexibility. RL has also revealed the pattern to follow for optimal daily scheduling of the storage tank, leading to the identification of Rule-Based control more efficient than the one used as benchmark.

Acknowledgements

ACKNOWLEDGMENTS

.....

Table of Contents

List of Tables

List of Figures

Acronyms

AC Alternative Current

AHU Air Handling Unit

AI Artificial Intelligence

ANN Artificial Neural Network

A3C Asynchronous Advantage Acotr-Critic

BCVTB Building Control Virtual Test Bed

BD Big Data

BEMS Building Energy Management System

BESS Battery Energy Storage System

CCS Carbon Capture and Storage

COP Coefficient Of Performance

DC Direct Current

DNN Depp Neural Network

DQN Deep Q-Network

DRL Deep Reinforcement Learning

HVAC Heating and Ventilation Air Conditioning

ML Machine Learning

MDP Markov Decision Process

IoT Internet of Things

MPC Model Predictive Control

nZEV nearly Zero Energy Building

PID Proportional Integrative Derivative

PV PhotoVoltaic

RBC Rule-Based Control

RL Reinforcement Learning

SAC Soft Actor-Critic

SC Self-Consumption

SoC State of Charge

SS Self-Sufficiency

TES Thermal Energy Storage

ToU Time of Use

VRES Variable Energy Resources

Chapter 1

Introduction

The energy sector has always been crucial for centuries, providing the basis for the economic and technological development, as seen during different industrial revolutions throughout history. The XX century highlighted how the economic growth is strictly related to the energy availability and how energy crisis can generate recessions, inequalities and contentions; indeed, the rapidly growing world energy use has already raised concerns over supply difficulties, exhaustion of energy resources and heavy environmental impacts like ozone layer depletion and climate changes. Nowadays, it is understood that energy production must be sustainable, to ensure stability in the long run and encourage prosperity to the human life. The current fossil fuel reserves give enough time to think about alternatives but it is not same for their impact on the global climate, which shows a much sooner deadline. The path has been revealed by many international organizations in the last decades, from the first United Nations Framework Convention on Climate Change in 1992 to the Paris Agreement in 2015 where the Parties committed to limit the temperature increase well below 2°C; also, the Climate-Energy Framework 2020 sets three key targets to cut 20% in greenhouse gas emissions (compared to 1990 levels), increase the EU renewables share by 20% and improve energy efficiency

by 20%, and lately, Europe has stated to be already on track to meet its greenhouse gas emissions reduction target for 2020, and has put forward a plan to further cut emissions by at least 55% by 2030. By 2050, Europe aims to become the world's first climate-neutral continent[1]. Also, with the Agenda 2030, the Sustainable Development Goals are defined for the energy sector, where it is expressed the needs for everyone to access to clean and affordable energy. Despite these efforts in Europe, nothing has really changed wordlwide, the atmospheric CO_2 concentration is still increasing and the targets set in 2015 seems out of reach unless a steep reversal of trend occurs [2].

The International Energy Agency has gathered frightening data on energy consumption trends. From 1990 to 2018, the growth of annual primary energy supply and CO_2 emissions is around 60%, at an average annual increase of 1.7% [3]. The trend of energy consumption over the past years helps forecasting how it will behave in the future and shows that new and more stringent policies must be issued by government and intergovernmental organizations, by intervening on the penetration of sustainable production, more efficient management strategies and technical improvements for distribution, storage and consumption technologies, along with CCS and carbon sinks, allowing negative CO_2 emissions. In any other case it will not be possible to achieve good performance in terms of global efficiency and so energy intensity, which is a major key indicator to understand effectiveness of strategies at macro-scale and to study the dependency between economy and energy consumption, that is growing at an astonishing pace. The U.S. Energy Information Administration provides energy forecasting in its International Energy Outlook 2019, where the reference case reports the growth of the world energy consumption to be nearly 50% by the end of 2050 with non-OECD countries accounting for around 87% of the increased amount, and among them, Asian countries are those contributing the most[4]. China and India have been among the world's

fastest-growing economies during much of the past decade, and they will remain primary contributors to future growth in world energy demand. Moreover, the IEO 2019 shows that the forecasting for non-OECD countries have higher uncertainties. Peréz-Lombard et al.[5] carried out the analysis of the trend of main world energy indicators between 1973 and 2004. The rate of population growth is well below the GDP, resulting in a considerable rise of per capita personal income, global wealth and energy consumption over the last 30 years. Electrical energy consumption has seen a dramatic increases of over two and a half times and scored a percentage increase of 18% in the final energy consumption in 2004 and this can be seen as important factor for final and primary energy intensities, which dropped because of the higher rate of growth of the GDP over the energy consumption increasing, resulting in an overall improvement of the global energy efficiency[5].

In this context, the present work is dedicated to the energy use in the building sector, which is gaining importance according to the historical trends. In 2010, buildings accounted for 32% of total global final energy use, divided in 24% for residential buildings and 8% for commercial ones[6][7]. In residential buildings, space heating dominated the consumption with a quote of 32% of the global consumption, followed by 29% for cooking, 24% for water heating, 9% for appliances, 4% for lighting, and 2% for cooling. Also in commercial buildings, space heating dominated the consumption with a quote of 33% of the total consumption, followed by 16% for lighting, 12% for water heating, 7% for cooling, and 32% for other equipment[6]. Moreover, the energy consumption in buildings is very heterogeneous across regions, differentiating by income levels, climate, and behaviour. This results in developed countries scoring up to 42 GJ/cap/yr, half of which used for space heating and fuelled for 73% by electricity and gas, while developing countries come up with 11 GJ/cap/yr, used primarily for cooking (47%) and fuelled with biomass (53%)[8]. The IPCC stated that GHG emissions from the building sector more

than doubled between 1970 and 2010, reaching a value around 10 Gt$CO_{2,eq.}$/y nowadays[9]. Also, the general believe is that building energy use will continue growing in the next 20 years, sustained by growth in population, increasing demand for building services and comfort levels, together with the rise in time spent inside buildings[3], and electricity is the fastest-growing source of energy use in the commercial sector, mostly due to space conditioning. An important driver is the global building floor area which is increasing at an annual average rate around 2.3%, supported by the growing population and increasing floor area per person[10]. The EIA analyses and forecasts future trends in building energy consumption. Energy use in the built environment will grow by 50% in the next 30 years, at an average rate of 1.3%. Many drivers such as economic and population growth in emerging economies will intensify needs for education and health, as well as public and private services, resulting in a strong energy consumption increase. Again, building energy consumption in non-OECD countries will increases at about 2% per year, about five times faster than in OECD countries, and non-OECD building energy consumption will surpass that of OECD countries by 2025.[4]. Heating, Ventilation and Air Conditioning (HVAC) systems constitute the major energy consumption in buildings with the percentage up to 60%[11]. In the USA, these systems represent more than 50% of the energy consumption in residential buildings, and in China, a sample of 30 buildings exposed a 68% of residential consumption in average[5][12]. Also, retro commissioning of existing building HVAC systems discovered the deficiency problems are mainly caused by control and operation[13]. Chiller plants are largest energy end-users in HVAC system, taking up more than 60% of the system whole energy consumption[14] and raise the problem of finding the Optimal Chiller Loading to enhance buildings energy efficiency. In Europe, data produced by the administration at national, regional or local levels is insufficient to efficiently plan future energy policies for buildings and to coordinate measures

to address each of the end uses. However, many sources show a significant increase in the use of air conditioning, especially in southern countries, probably due to the global warming, creating serious supply difficulties during peak load periods. Analysis by sectors, as those produced by the EIA for residential and commercial buildings should be funded by governments, so that a comprehensive database of the building stock and energy parameters can be the basis for future planning. For these reasons, energy efficiency in buildings is now a prime objective for energy policies at regional, national and international levels. Since 2013, government and regulators are developing strategies at a large scale concerning either home appliances, water, and space heating in residential buildings and space heating and other miscellaneous equipment in non-residential buildings[15]. It is only recently that BRICS countries have issued more stringent efficiency standards for appliances and equipment, as well as stricter building codes. The building envelope has gained attention as well. In 2013, IEA published a report stating that US, EU, and Russia should primarily intervene on high performance envelopes in the cold areas, where the building energy demand could be potentially reduced by 33%[6]. Behavioral changes should be taken into account, but it is difficult to carry out any accurate analysis. It is generally assumed that behavioral changes could save between 10% and 30% in heating, up to 50% in cooling and up to 70% in lighting [9].

At this time, many achievements have been made to fulfill the energy-efficiency requirements for equipment in buildings, by guaranteeing the operative needs and being environmentally friendly, but further policies are required to drop equipment costs down[16]. One of the main instrument issued in Europe in the building sector is the Energy Performance of Building Directive (EPBD), which sets sets the standards for new and renovated buildings for each EU Member States (MS). The measure is is expressed by the Directive 2010/31/EU at Art. 9 where it is indicated that EU Member States (MS) must ensure that by 2021 all new buildings,

and already by 2019 all new public buildings, are nearly Zero Energy Buildings (nZEB), and also, it encourage MS to draft plans and best practices as regards the cost-effective transition from existing stock to highly efficient buildings[17], but the pace of renovation is much slower than needed, and yet, the developed countries are going to face more than an hurdle in replacing the old existing building stock. For instance, the US and EU have an average replacing pace of 2% and 3%, respectively, which means that the energy consumption must be reduced by promoting both high-performance buildings and retrofitting practice[18].

Recently, Building Energy Management Systems (BEMS) are developing and are being empowered more and more thanks to the implementation of Internet of Things (IoT) and Cloud Computing for Big Data gathering to support building managers and proprietors and to improve capacity, cost-viability, adaptability, accessibility, effectiveness, toughness, and dependability, in new and existing buildings, both residential buildings and non-residential[19][20]. IoT enables smart things to communicate with each other, and incorporate real-world data and knowledge into the digital world. Smart devices with sensing and interaction capabilities, as well as recognition technologies, make it possible to collect far more knowledge about the real world than ever before. This wireless communication has broadened thanks to sensors for various applications like smart buildings, smart cities, smart healthcare, and the smart industry[21]. Big Data technology is referred to the huge amount of data collected by these sensors which is also allowing to have into operable insights and more accurate predictions[22]. This new concept has expanded the possibility for maintenance and efficient management of building energy systems. In this new context, consumers play an active role in balancing the grid operation by changing and possibly compromising on their current consumption patterns to enhance the building flexibility. Consumers will be able to have insight and control of their Electric and Electronic Equipment (EEE) in an effective and efficient manner.

DSR services are so enabled and potentially provided to system operators, who need information on supply of flexibility available when the whole energy system is undergoing high electricity generation, peak load or generally the system is off balance[16]. However, the incremental initial costs of green (energy efficient) buildings has been reported as a significant barriers to high-performance buildings with the ultimate goal of achieving net-zero energy by design teams and building owners[23][24].

Deep research is on going for the systematic characterization of energy use in buildings. Different building end-use have different energy use profiles, and characterization of the major contributors and their energy use is needed. Residential and commercial buildings features mainly energy consumption associated to the comfort level of the occupants, while industrial buildings consumption is primarily due to the operation of industrial machinery and infrastructures dedicated to production processes[21]. Based on the type of building, different management strategies can be used to achieve energy savings, and there are uncountable types of building across different final end-uses and configurations, so that buildings clustering is typically used for aggregations based on load pattern, end-use energy and occupancy schedule rather than technical parameters. Buildings such as residential, education, office, healthcare, and industrial are emerging as critical consumers in energy consumption[19], thus making energy flexibility the key point for smart buildings, that need to be able to manage high-complexity dynamic system considering occupants behaviour, storage systems, renewable on-site generation, electric vehicles and Demand Response programs through a continuous information exchange[25]. As buildings energy system becomes more complex, more sophisticated control strategies have to be implemented for managing energy fluxes, and computational cost for modeling increases as well. Literature provides examples of smart control for lighting, heating, cooling and electrical appliances. This work is focused on the

cooling system control.

In the field of HVAC systems, the most common techniques are ON/OFF control, P, PI or PID control, and for this reason, they are typically referred to as classical control strategies. The rule-based prescriptive approach guarantees occupant comfort by maintaining a comfort range. Additionally, it is possible to reduce energy consumption and carbon emissions by adjusting the setpoints based on heuristic rules. ASHRAE Guideline 36 summarized those rules , which could represent the state of the art of this approach adopted by industry[26]. An on/off controller regulates the ON/OFF state of a component in order to keep a certain value within a threshold, but they can not deal with dynamic systems[25]. P, PI and PID controllers modulate a controlled variable by using error dynamics as long as the operating conditions do not vary from the tuning conditions[27]. They also require a laborious tuning of parameters. These control strategies are very simple and effective, but not optimal. Mainly due to the lack of predictive information; indeed, it is not possible to take into account the day-after prediction and anticipate the behaviour of the system. Also, the control sequence is fixed and predetermined, thus it is not customized to a specific building and climate condition[26].

To overcome classical control limitations, Model Predictive Control has established in the building control research community and it is proven that this control method can achieve energy savings while maintaining or even improving thermal comfort in buildings. This techniques relies on the modeling of the energy system, the prediction of disturbances and finally solves the control problem. Since it was initially proposed in the 1970s in the chemical and petrochemical industries, MPC has been successfully applied in many fields[28].

As regarding to building control, MPC has been applied to radiant ceiling and floor heating [29][30], intermittent heating [31] and ventilation [32], and to optimize cold water thermal storage systems[33]. Eventually, MPC proved to be

effective not only during simulation, but also in real-time[34]. MPC deals very well with non-linear time-varying disturbances and are able to predict future evolution therefore they find many application in HVAC operation. According to Afram and Janabi-Sharifi[27] MPC can explicitly handle disturbances, uncertainties and constraints and include predictions of occupant behaviour, equipment use and weather forecasting. Moreover, it is possible to consider a multi-objective cost function as well as deploing MPC both at supervisory and local control. Nevertheless, the model complexity can become unsustainable in terms of computational power and effort to build the model, also, they are very sensitive to any changes in the environment. Again, Afram and Janabi-Sharifi stated that the integration in HVAC systems may be difficult or impractical, due to the specification of many parameters, resulting in a labor-intensive process and required expertise to use.
In the building sector, this results in a very low flexibility given the diversity of the existing stock, where building and its energy systems are unique, so it is difficult to generalize a standard building energy model for various buildings. For this reason, the spread of MPC has recently slowed down despite the promising results and allowed soft control being considered as a possible solution[35].
Soft control systems are based on fuzzy logic, neural networks or genetic algorithms and are usually applied for supervisory control[25]. Soft controllers are not very common in real building applications since their accuracy often relies on the quantity and quality of available data points, hence Big Data is an enabling technology for this kind of controller. ANN-based control systems are trained thanks to sets of historical data, which must be not only large enough to cover a wide range of operating conditions but also they must ensure quality in terms of accuracy and time sparsity. Similarly, fuzzy logic controllers require an extensive knowledge of the building operation under different conditions, but the control strategy is represented by a set of rules which defines the operational phases according to the

information fed to the controller. Among soft control, Reinforcement Learning is gaining attention, although research in the building control field has seen MPC as main focus, Reinforcement Learning based controllers have shown remarkable progress in many difficult and previously unsolvable domains[36][37][38]. RL belongs to the Machine Learning framework which has demonstrated its potential to enhance building performance at many stages of the building lifecycle[39], thanks to BG, more powerful computing and algorithm advancement. As follow, it is introduced the general framework of RL, then in the next chapter, it is provided a detailed description of the algorithm and the modification of its main component. RL based control can be classified into two further subclasses, each one with its peculiarity: model-free and model-based RL. Model-free RL can be seen as the counter-part of MPC ; indeed, model free RL does not have any knowledge of the environment, rather it learns the behaviour by directly interacting with it. Its working principle can be summarized as follows: it observes the system state, choose an actions and observes the next state and the reward it obtains from the environment. Its task is to maximize this reward stream over time which includes also future rewards. Real world control problems often are trying to achieve multiple objectives like energy consumption, energy cost or users comfort, which means that defining the controller's reward function must take into account the competitiveness of these objectives by properly weighting. The main advantage of Model-free RL controllers is their lower computational cost compared to MPC, due to the absence of a model; infact it is data-driven, and it helps avoiding the work of developing and calibrating a detailed model, as it is for MPC. During operation, the RL agent not only executes the optimal control action but also updates its policy which can make it robust under dynamically changing environment. The robustness and the smaller computation cost as well as the lower efforts in building a model can make RL better suited to many practical implementation when compared with

classical MPC controllers, given that enough data are available for training. On the other hand, model-free controllers have significant limitations at the current stage. Mainly the curse of dimensionality, the need for a huge amount of data, that are lacking for the most of the case, before it can discover a nearly optimal policy, and also the 'sample complexity', which is usually much lower than their model-based counterparts. Furthermore, during the training phase, exploratory actions are taken by the RL agent to discover the goodness of different trajectories, but such actions can directly lead to occupant discomfort or to excessively high energy cost that need to be accounted as cost of implementation. These last issues could be worked around by constructing virtual simulation environment and using it to train the agent, but this solution increases the computational cost, brings in the efforts of model implementation and the associated loss of accuracy. The Open AI Gym is an online library that allows to build training environment for RL algorithm. Model-based RL has been developed to overcome these limitations, and it is used extensively in robotics and other disciplines where decision making needs to happen in real-time with limited and noisy sensing data. These algorithms can perform as well as MPC while also offering the potential to greatly reduce computational complexity. Model-based RL learns the transition probabilities from a state given a certain action, for each state. For this reason, they have many similarity with data-driven MPC, but they differ in the way of optimizing the control strategy, mainly due to the explicit exploratory strategy where the controller is encouraged to explore the state space better to discover potentially rewarding strategies, and the use of policy-side learning to speed-up computation. Policy-side learning means that the model-based RL controller learns from solving these optimizations over time and does not need to repetitively optimize when incurs in similar states. The biggest advantage of model-based RL is its compatibility with existing MPC controllers. By offering similar or better performance at (asymptotically) reduced

computational loads, it can contribute to the next generation of controllers. The main drawbacks concern the additional system complexity which can potentially increase the likelihood of a failure. In addition, depending on the size of the state-space, policy-side learning can be a formidable undertaking requiring substantial amounts of computational resources before convergence. Furthermore, exploratory steps can likewise improve the asymptotic performance of the controller while risk of lost user comfort increases, as with model-free controllers. Definitely, data-driven models give added-value to the huge amount of data that is collected everyday and can substitute complex physical models with pure mathematical relations, speeding up the construction of the model, reducing computational cost and lowering the number of input, but more data need to be collected. Today, new enabling technologies for data management have allowed a bi-directional flow of information, but the smart buildings involves higher investment costs that need to be justified by a strong reduction in energy cost, which is not happening yet. At this time, costs are the major barrier for penetration of the smart technologies, and policies must help both companies and users to establish this new framework, in order to increase the amount of data available and their quality. Technological advancement should be assisted by advanced methodologies for building control, but neither MPC, neither RL are competitive with classical control at the current stage because of their low scalability. Once RL capabilities are going to be more clear, as well as its application, which is the aim of this work, research should address the problem of the length of the learning phase. This is the main barrier preventing from real-time implementation. In this sense, a few solutions are to be addressed more consistently in order to understand which is the path to follow. Transfer learning seems a feasible approach, where RL agent is trained in advance and then it is dynamically deployed on many similar buildings, thus obtaining the scalability factor. This is achieved by developing open-source training environment

where everyone can pre-train the RL agent. Also, the expert knowledge could be encoded and achieve the reduction of the training phase, but this is not enabled by some of the RL algorithms.

1.1 Previous works on RL algorithms

At the current stage, literature provide examples of various RL methodologies but real-time applications are pretty much concentrated in the academic context. ANN's have massively spread in the last decade of the XX^{th} century and established as the basis for data-driven models in energy forecasting[40] thanks to the more sophisticated computing technologies, whereas the reinforcement learning has only recently become popular for managing HVAC operation and considered as a promising technique in the energy system control research field. Hong et al.[41] reviewed the field research where the efforts have been focused. The main applications concern the building design, operation and control, with an increasing number of yearly publications over the last decade. In this section, it is presented a short review on the building control RL algorithm. The most popular RL algorithm for building control is the Q-Learning methodology, where the agent learns the value of each state-action pair.

The first approach considered discrete state-action space with tabular Q-value. Liu and Henze[42] in 2006 used tabular Q-learning to control the temperature set point and the operation of thermal storage. May and Ross[43] achieved to control the window opening state choosing between on and off while improving the occupant's comfort. The same action space was considered by An et al.[44], who proposed a RL approach to reduce indoor PM2.5 concentrations in naturally ventilated buildings without air cleaners. A DQN algorithm was trained in a specific naturally ventilated apartment for a one-month period, and successively deployed. The study concluded that the RL control works better in both virtual

and real apartment than the baseline I/O ratio algorithm. The average indoor PM2.5 concentration was reduced by 12.80% in the course of a year for a virtual environment and by 9.11% for the real one. Moreover, RL reduced by 7.40% the PM2.5 concentratrion when compared with real window behavior. Qiu et al.[45] applied the Tabular Q-Learning to improve the global COP of the cooling water system of the HVAC system serving a underground station in Guangzhou. Then, it was compared to three other control systems (baseline controller, local feedback controller and model-based controller). The results showed the Model-free control could save 11% of the energy, which is more than 7% in local feedback controller but less than 14% of model-based.

As the number of state-action pair increases due to the dimensionality or to continuous values. Yu and Dexter[46] integrated fuzzy rules and Q-learning to control an HVAC set point, while Zhou et al. [47] achieved to manage a smart grid. Differently, it can be used a linear function to approximate the Q-value, as in Dalamagkidis et al.[48] to control the HVAC operation. Another way to approximate the Q-function is to use the fitted Q-iteration approach. This is used in Leurs et al.[49], who attained the peak shaving of the maximum feed-in power of a PV system into the grid by controlling an HVAC system. Also, Ruelens et al. used this type of controller for a heat pump's operation [50] and De Somer et al.[51] for a domestic hot water heater. Recently, Deep Learning has enhanced the potential of function approximators by exploiting Deep Neural Network. An example is provided by Vázquez Canteli et al.[52] who applied it for the control of a thermal storage operation. A different application is proposed by Brandi et al.[53], in which an algorithm based on Deep Q-Network has been used to control the supply water temperature of the boiler serving the radiant heating system installed in an office building. In this work, an online network and a target network are initialized, the first one is constantly updated and directly used in the interaction with the environment; the second

one, called target network, is updated after N iterations and used to predict target values. In this context, a static and dynamic deployment of the DRL controller is performed, and a heating energy saving ranging between 5 and 12% is obtained with enhanced indoor temperature control with both deployment. The same approach is used by Yoon and Moon[54] to minimize the energy consumption. A performance based thermal comfort control using Double Deep Q-Network allowed to reduce by 32.2% and 12.4% the energy consumption associated with the Variable Refrigerant Flow (VRF) and humidifier system, within acceptable Predicted Mean Vote limits. The control action were the ON/OFF status of the humidifier, the temperature set-point and the VRF airflow rate. Also, Ding et al.[55] adopted a double deep Q-Learning named OCTOPUS, employing a novel Deep Reinforcement Learning (DRL) framework that uses a data-driven approach to find the optimal control sequences of all building's subsystems, is used to minimize the energy used in heating/cooling coils, the electricity used in the water pumps and flow fans in the HVAC system, electricity used by the lights, and the electricity used by the motors to adjust the blinds and windows. In addition to the minimization of energy, it is requested maintaining the human comfort metrics within a particular range. Through extensive simulations it is demonstrated that OCTOPUS can achieve 14.26% and 8.1% energy savings compared with the state-of-the art rule-based method, while maintaining human comfort within a desired range. Gupta et al.[56] simulated a multi-building scenario under different assumptions, and compared a Deep Q-Network (DQN) with a classical thermostat control. The thermostat control has a +/- 3°C deadband around the optimal indoor temperature, while the DQN acts directly on the ON/OFF status of a heater. The outcome showed the RL algorithm outperforms the thermostat-based controller by improving both thermal comfort, as deviation from optimal indoor temperature, and heating energy consumption. Time-varying electricity price profiles have been investigated rarely,

Jiang et al.[57] have recently applied a DQN with action processor in order to enable planning one or more day ahead, under Time of Use tariff and demand charges conditions. The objective function considered energy cost and a discomfort penalty, and it was subjected to shaping technique to overcome the issue of reward sparsity caused by the demand charge. A single-zone building was simulated, and it was demnostrated that the customized DQN outperforms the baseline, saving nearly 6% with demand charges, 8% without demand charges of the total energy cost.

A different approach is to have the policy function parameterized but high variance and difficult convergenge make it difficult to implement in the building control[41], even though, the literature reports some contributions. Chen et al.[58], used a Proximal Policy Optimization (PPO) RL algorithm to control supply airflow rate of an AHU and supply water temperature of floor heating, achieving 7%–17% energy conservation compared with the benchmark. Azuatalam et al.[59] used a modified version, namely PPO-Clip to ensure minor deviations between new and old policies. The agent chooses the zone temperature set-points of a whole commercial building for several Demand Response (DR) scenarios, where it was able to modify the power consumption according to the DR signals and to guarantee acceptable comfort level.

The need of encoding expert knowledge or pre-training in the actor network has been met by developing an actor-critic network, where both Q-value and policy are parameterized. In this way, it is allowed to store and reuse the weights stored in the actor network, rather than initializing randomly[60]. The main drawback is the increased computation cost[41], so that actor-critic is not very popular in the building control field. An actor-critic neural network approach was applied to adjust the signal of a local control for HVAC control in 2008 by Du and Fei [61]. This study reports significant improvements from a combined PID actor-critic

learning approach than a stand-alone PID controller. Actor-critic has been studied again by Fuselli et al.[62] and Wei et al.[63] for energy storage control, by Al-Jabery et al.[64] for domestic hot water control, and by Bahrami et al.[65] to optimize the scheduling of smart home appliances. The literature provides examples on more actor-critic algorithm, Zhang et al.[66] applied an RL control type Asynchronous Advantage Actor-Critic (A3C) in a water-based Radiant Heating System, where the hot water pipes are integrated into window mullions. The goal is to reduce the energy consumption of the system while respecting the internal comfort of the occupants. The control system, in this case, operated on the mullion system supply water temperature set-point. The same objective was achieved by Zou et al.[67] using a Deep Deterministic Policy Gradient algorithm trained thanks to a Long-Short-Term-Memory (LSTM) networks approximating real-world HVAC operations of three AHU's system and controlling fan speed, heating valve status and damper position. In Park and Nagy[68], it is presented a Reinforcement Learning based Occupant-Centric Controller (OCC) for thermostats, called HVACLearn. The agent learns the unique occupant behaviour and indoor environments and monitoring indoor air temperature, occupancy, and thermal vote. The objective is to find the optimal trade-off between the users comfort and energy consumption, by adjusting the thermostat set-points. Authors simulated HVACLearn control in a single occupant office with occupant behaviour models. HVACLearn control is so compared to the baseline, and it was able to reduce the number of button presses (too hot) significantly, while consuming the same or less cooling energy.
Biemann et al.[69] implemented different RL algorithms, which reduced energy consumption with respect to model-based controllers by more than 13%. Particularly, SAC algorithm showed clear improvements the first year and needed up to ten times less data. Its high data efficiency and stability was considered to favour real world applications.

Multi-agent algorithm have been explored as well. Yu et al.[70] presents a Multi-Agent DRL (MADRL) called Multi-actor attention critic (MAAC), in order to minimize HVAC energy cost in a multi-zone commercial building under dynamic prices, with the consideration of random zone occupancy, thermal comfort and indoor air quality comfort in the absence of building thermal dynamics models. To be specific, air supply rate in each zone and the damper position in the air handling unit are jointly determined to minimize the long-term HVAC energy cost while maintaining comfortable temperature and CO2 concentration ranges. For encouraging exploration, Soft actor-critic (SAC) method is used. The simulation results showed the effectiveness, robustness, and scalability of the proposed algorithm. Nagarathinam et al.[71] consider the optimal control problem of minimizing the building HVAC energy subject to meeting the comfort constraints by dynamically setting both the building and chiller set-points. In this frame, it is presented MARCO (Multi-Agent Reinforcement learning Control) for HVAC system. MARCO is based on Double Deep Q-Network algorithm and uses separated DRL agents that control both the AHU's and chillers to jointly optimize HVAC operations. Authors train and deploy the agent in real configurations and it is showed that MARCO learned the optimal policy in a two-agent setting with single AHU and single-chiller. MARCO not only improved comfort but also reduced the energy by 17% over a baseline that used seasonal variations in set-points.
Further research have been addressing the possibility of shortening the training phase, which is a huge hurdle for the spread of RL. R. Jia et al.[72] showed how to implement the expert knowledge on a DRL algorithm through "experience replay" or "expert policy guidance, in order to reduce the length of the training phase, but also, proposed to stabilize the learning process by penalizing the erratic behavior. Vázquez-Canteli et al.[73] built a training environment based on the OpenAI Gym library, in order to allow researchers to implement, share, replicate, and compare

their implementations of reinforcement learning for Demand Response applications. It is a customizable and modular framework where researchers can implement storage technologies, energy generation, or energy-consuming system, according to their purposes. Pinto et al.[74] made use of this training environment for a single-agent RL centralised controller to flatten the cluster load profile while optimizing energy consumption of each building. A SAC algorithm was used to manage 8 thermal storages of a cluster of four buildings equipped with different energy systems. The coordinated approach was compared against a manually optimised Rule-Based control for single buildings. Operational costs dropped down about 4%, while the peak demand was reduced by 12%. But mainly, the coordinated energy management allowed to reduce the average daily peak and average peak-to-average ratio by 10 and 6%, respectively.

The training phase could be reduced by means of transfer learning as well. Deng et al.[75] built and validated an RL occupant behavior model for an office building and transferred it to other buildings. Transfer learning was successfully carried out between commercial buildings with different HVAC control systems, and from office buildings to residential buildings.

1.2 Contribution from this work

Previous works have proven the RL based algorithms can provide nearly optimal policies for energy systems control. Nevertheless, further research are required to find advanced control strategies which have to guarantee adaptability to different environment, flexibility to the energy system and applicability in real operation. In this work, a DRL based on the Soft Actor-Critic algorithm is used to manage thermal and electrical energy storage in a multi-energy building during the cooling season. The energy system is made up by an HVAC to serve the thermal zones, an electric chiller, a cold-water storage tank, a PV module and a Li-ion battery. The

thermal side of the model is simulated through EnergyPlus, while the electrical side is simulated through Python. The simulation environment is built thanks to the OpenAI gym library. This work aims at describing how an advanced control strategy affects the design of the thermal and electrical storage. The control problem is defined by a control action which is the tank operation and a objective function, which is the energy cost. A Time Of Use tariff with 3 fares is used to compute the electricity price. Two control strategies are compared, the benchmark which is a RBC for TES with temperature set-points for charging during low price hours, and the SAC algorithm, which makes use of its agent to choose on the scheduling of the cold-water storage tank, so that the building load can be increased or decreased, hence the BESS operation is influenced as well. Both the strategies, manages the BESS through the same RBC.

The idea is to use a smart strategy in order to make a better use of the storage technologies, thus allowing lower size equipment to be effective and avoid unnecessary investment costs. Particularly, the analysis highlights how the definition of the control strategy itself is part of the design of an energy system. The proposed strategy is expected to satisfy the thermal and electrical load by means of lower size equipment, as well as reducing the total energy cost without any further effect on user comfort. Also, this work provides an overview of the advantage of adopting smart control strategy in the perspective of the compelling need for enhancing building flexibility and pushing toward nZEB. This means that RL should prove to be very effective at managing energy storage, and improve their profitability at each size. Mainly, BESS are crucial components for building energy systems, and it is urgent to find a way of increasing self-consumption level without scaling up. In this context, this work shows how RL can play a role in achieving those objectives. The chapter 2 introduces the RL framework and its theoretical foundations, along with the improvements that have been made during the last years. Chapter 3

describes the case study and formulates the control problem, as well as describing the two control strategies. In chapter 4, the RL algorithm design is presented by defining the state space, the action space and the reward function and the hyper-parameters. The simulation environment is described and so is the experimental setup, which defines the configurations for the storage systems. Chapter 5 shows the results throug plots and tables, and then, these are discussed in the next section. Finally, chapter 6 provides the recap of the work, analyze the main implications and open the questions on how to improve the study and on future contribution.

Chapter 2

Fundamentals of Reinforcement Learning

Reinforcement Learning belongs to the Machine Learning discipline, together with supervised and unsupervised learning, as seen in Fig. 2.1. The main difference lies in the ability to assign a score for the output, which tells how valuable a state is[76]. The algorithm has a learning agent which chooses the actions to be carried out in order to achieve a certain objective, by interacting with the environment where it is implemented. Decisions are made at each control time step and depends on the current state of the system, and the way the agent reacts to the external disturbances defines the so called policy. RL is applied to problems that can be divided into two categories [48]: i) Episodic problems, that have one or more terminal states. An episode is repeated over, during which the agent is trained, meaning it investigates all possible combinations of states and rewards. When the agent gets to the terminal state, the episode ends, and the environment is reset to the initial state for a new episode to start. ii) Continual problems do not end, and they continue indefinitely. RL refers to the framework referred to the Markov

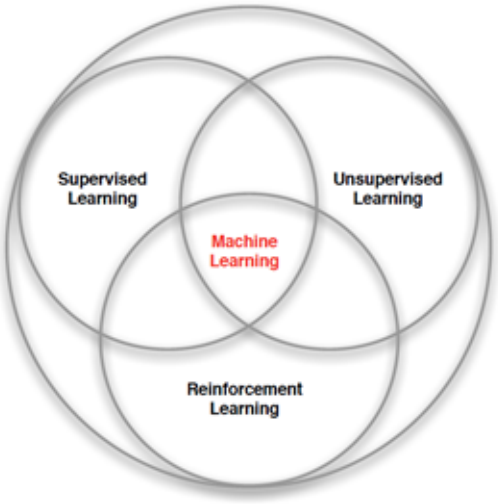

Figure 2.1: Machine Learning typologies

Decision Process (MDP), according to which both the reward and the probability of transition between the previous and the next state depends only on the current state and the action chosen. MDP predicts the following state and the expected reward exploiting the information available at the last time step and rejecting all those information from the past experience. MDP formalizes the information exchanged between agent and environment mathematically, and indicates the main elements that made up the RL agent[77]):

- State space (s ∈ S), all possible states of the environment considered. The definition of the state space may be subjected to a sensitivity analysis to find out which are the most important variables that come into play. Also, if an unnecessary state is chosen, the RL agent suffers from the curse of dimensionality [26].

- Action space (a ∈ A), the set of possible actions that can be selected by the agent at each timestep.

- Reward (r), a scalar value that indicates how much the objective is accomplished and it is sent back from the environment after performing the action chosen by the control agent.

- Policy (π), it represented by the mapping of the actions probability distribution for each state. The agent's objective is precisely to acquire an optimal policy.

- Transition probability distribution, which is present in model-base RL and specifies the probability of the environment to end up in a certain state given the current state s and the action a.

For each control time-step, the agent will perform a particular action, and the environment issues both the scalar reward and the information about the state to the agent. Tuples are mathematical elements representing the current control time-step. It is a vector that contains within it four elements: state, action and reward at the current time-step and state at the next time-step. The agent takes advantage from two value functions to define the policy, these are the state-value function and action-value function, which provide the goodness of states and actions. The state-value function represents the expected reward given by the agent when starting from a state s, following a specific control policy π [48]. The following equation expresses it:

$$v_\pi(s) = E[r_{t+1} + \gamma v_\pi(s')|S_t = s, S_{t+1} = s'] \tag{2.1}$$

The action-value function represents the expected reward given by the agent when it chooses an action a, starting from a state s, following a specific control policy π. The following equation expresses it:

$$q_\pi(s, a) = E[r_{t+1} + \gamma q_\pi(s', a')|S_t = s, A_t = a] \tag{2.2}$$

where γ is called the discount factor, it is in the range between 0 and 1, and defines the importance of future rewards with respect to the present reward. If $\gamma = 0$, the agent will consider only current reward, neglecting future reward. When $\gamma = 1$, current and future rewards have the same importance. These two functions are updated online during the training phase of the agent, so they depend heavily on the experience gained. The RL agent is trained through a trial-and-error approach: this means that it explores different trajectories (i.e. policies), receive a feedback from the environment, and it is continuously updated to improve them. This is what is referred to as on-policy learning, i.e. the policy output of the controller is being carried out by the environment. The agent can also learn from other policies that have already been implemented in an environment, in this case, it is called off-policy learning. Some value-based algorithms use this methodology, especially for its greater flexibility than on-policy learning. However, off-policy learning suffers from a lower inclination to explore action space, if compared to its counterpart. To overcome this problem, a large amount of measured data should be available, but using only measured data may be inadequate [26]. So, as in our case, simulated virtual environments can be created and used to train the RL agent. To do this, it is advisable to create an interface between energy simulation and control platforms, such as EnergyPlus and Python. The function described above can be dealt by RL in different ways, so that the algorithms are becoming more and more complex and the range of applicability is increasing. The simplest way to compute the Q-values is to store the expected returns in a look-up table, called Q-table, which is updated according to the Bellman's equation [78].

$$Q(s,a) = Q(s,a) + \alpha[r_t + \gamma max_{a'} Q(s,a) - Q(s',a')] \tag{2.3}$$

With α [0,1] named as the learning rate and establishing how much new knowledge overrides old knowledge. $\alpha = 1$ means that new knowledge completely overrides the old one. As the dimensionality of the space of actions or states increases, or they have continuous values, the memory storage and computation time required to update the Q-table increase, so Q-Learning becomes inadequate [79]. In this case, an alternative to the tabular form of Q-Learning has been elaborated, it may be useful to use a Deep Neural Networks (DNNs) as function approximators. The main element of a neural system is the neuron, composed of a cellular body and an axon that sends the output response to the next layer. The structure is a dendritic tree that connects the neurons of different layers. The topology of a DNN is based on multiple layers of neurons. Typically, a neuron is a non-linear transformation of a linear sum of its inputs and it can be expressed by the following equation:

$$output = f_{activation}\left(\sum_{\#neurons} input_i + bias\right) \tag{2.4}$$

DNNs are composed of input and output layers, but between them are hidden layers that take input from the previous layer and perform a mathematical operation. Fig. 2.2 shows the graphical representation of a neural network. By inserting DQNs into the Q-Learning results Deep Q-Learning (also called as Deep Q-Network). The Q-values will be indicated with the following formula, taken from Nair et al.[80]:

$$Q(s, a) = Q(s, a; \theta) \tag{2.5}$$

The equation represents the Q-network, where the term θ parameterizes the Q-value function: it indicates the weights of the network. The input layer has as many neurons as the number of variables belonging to the state space, while the number of neurons in the output layer corresponds to the dimension of the action

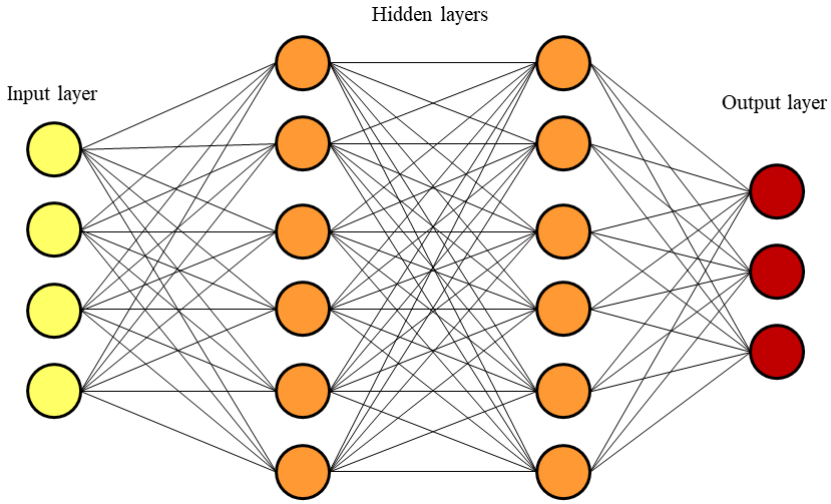

Figure 2.2: ANN basic structure

space. The whole network can be seen as simply a function that expresses the relationship between states and Q-value, for each action, in order to find the optimal policy. It is good to remember that this last parameter is not known a priori and is obtained during the training process as already explained in the previous paragraph on Q-Learning, and updated according to Bellman's equation. The literature provides example where it is shown how the use of two DQNs, namely online network and target network, improve the performance of the learning algorithm. The main characteristic of this technique is the ability to counteract the overestimation of the Q-values that may lead to a non-optimal outcome when using a single DQN. This target network is a duplicated of the other one, but it is synchronized every τ steps (an arbitrary number), and it is used to calculate the target Q-values for expectation [81]. Instead, the online network is the one used to interact with the environment and updated with Bellman's equation. The use of a replay memory, helps storing the tuples referred to the previous experiences of the agent: this allows, if necessary, to reuse them and go beyond the problem of related observations. At the same time, the optimization process is carried out. This

work makes use of a target network and a replay buffer, but with a discrete action space to explore. Deep Reinforcement Learning algorithms face problems related to high sample complexity, so that simple tasks could require a huge number of data collection steps: this leads to poor sample efficiency due to on-policy learning. It is necessary to try to switch to off-policy algorithm. Moreover, they suffer from dependence on the chosen values of hyperparameters, like discount factor, learning rates, exploration constants and other. These two obstacles make it challenging to apply these control algorithms to real-cases. To try to overcome these obstacles, the Soft Actor-Critic (SAC), an off-policy algorithm based on the maximum entropy RL framework, was recently introduced by Haarnoja et al.[82]. While most existing model-free works make use of discrete action space, SAC easily allows working with a continuous one. This algorithm aims to maximize a target function composed not only of the term expected reward but also of an entropy term. This last term, is what expresses the attitude of our agent in the choice of random actions. It also has dual importance, as it ensures that the agent is explicitly pushed towards the exploration of new policies and at the same time avoids that it transposes lousy policy. SAC uses the entropy terms to represents the trade-off between exploration and exploitation. Exploration is defined as the phase in which the agent is neglecting his real goal of maximizing the reward and samples actions within a new set composed by those not yet selected. This is needed in order to avoid trapping in local minima. By exploitation, it is intended that phase in which the agent chooses within the previously selected actions, the one that allows him to get the higher rewards considering the knowledge it has acquired. A right control agent must try to optimize the compromise between these two stages. The current state of the art sees applications like those in the field of robotics, and recently it is increased its application in energy building context. As proposed in Haarnoja et

al.[82], the Soft Actor-Critic algorithm incorporates three key ingredients: an Actor-Critic architecture with separate policy and value function networks, an off-policy formulation that enables reuse of previously collected data for sample efficiency, and entropy maximization to encourage stability and exploration. The SAC presented around the last months of 2018 suffered from dependence on hyperparameter temperature, therefore in the latest version proposed in Haarnoja et al.[82], it is devised an automatic gradient-based temperature tuning method that adjusts the expected entropy over the visited states to match a target value. Soft actor-critic is based on the maximum entropy reinforcement learning framework, in which the objective is maximize both expected reward and entropy. It could be seen as an extension of standard RL objective. The maximum entropy objective requires an optimal policy like this:

$$\pi^* = argmax_\pi \sum_t \gamma([E_{s_t,a_t}[r(s_t,a_t) + \alpha H(\pi(\cdot|s_t))]]) \tag{2.6}$$

with α temperature parameter, that indicates the importance of the entropy term compared to the reward one, it also indicates the stochasticity of the optimal policy. Generally α is zero when considering conventional reinforcement learning algorithms. It is convenient introducing a discount factor γ to ensure that the sum of expected reward and entropies is finite. The SAC is derived from a variant of the maximum entropy framework, called Soft Policy Iteration, which is not presented here. The state-value function for SAC algorithm can be written as follow:

$$V(s_t) = E_{a_t \sim \pi}[Q(s_t,a_t) - \alpha log(\pi(a_t|s_t))] \tag{2.7}$$

When the observation space is continuous, the optimization of the soft q-function is performed by minimizing the soft Bellman residual:

$$J_Q(\theta) = E_{(s_t,a_t)\sim D}[\frac{1}{2}(Q_\theta(s_t,a_t) - (r(s_t,a_t) + \gamma E_{s_{t+1}\sim p(s_t,a_t)}[V_{\bar{\theta}}(s_{t+1}])))^2] \tag{2.8}$$

where D is the replay buffer and $V_{\bar{\theta}}s_{t+1}$ is estimated thanks to the target network and a Monte-Carlo estimate of Eq. 2.7. Also, it is needed to reparameterize the policy, whose losses can be expressed by Eq. 2.8:

$$J_\pi(\phi) = E_{s_t \sim D, e_t \sim N}[\alpha log(\pi_\phi(f_\phi(\epsilon_t; s_t)|s_t)) - Q_\theta(s_t, f_\phi(\epsilon_t; s_t))] \tag{2.9}$$

The reparameterization trick allows overcoming the problem of backpropagating errors in the normal way.

In the first version of SAC, the temperature parameter was fixed and then considered as an hyper-parameter, so its choice had an important influence on the agent's behaviour. To avoid this problem, in the next SAC update was introduced the possibility of making alpha as an update-able parameter. In particular, it is updated by taking the gradient of the objective function below:

$$J_\alpha = E[-\alpha \ln \pi_t(a_t|s_t; \alpha) - \alpha \hat{H}] \tag{2.10}$$

where $\hat{H}$ represents the desired minimum entropy, set to a zero vector. This SAC latest version improves both performances and the stability of the algorithm, and it is not implemented for this thesis work. The SAC policy is expressed as a Gaussian distribution which actions are sampled from, and optimized using approximate dynamic programming [82]. In conclusion, this algorithm is particularly useful under a changing environment or when agent's knowledge of the environment changes [83], also it becomes necessary when the action space is continuous.

In this work, a SAC for discrete action is chosen. Generally, a discrete action space involves a faster convergence since there are less state-action pair to explore, and mostly important, reduces the length of the training phase, which is the main drawback of using RL algorithms.

P. Christodoulou [84] derived a discrete version of the SAC described by Haarnoja et al.. The difference is that the policy $\pi(a_t|s_t)$ is not anymore a represented by

a probability density function but it has finite value, oppositely, the derivatives of the loss functions still hold. The following changes are also proposed by P. Chistodoulou:

- The soft Q-function outputs the Q-value of each possible action rather than simply the action provided as an input, i.e. our Q function moves from Q: S x A $\rightarrow$ R to Q: S $\rightarrow R^{|A|}$. This was not possible before when there were infinitely many possible actions we could take;

- The policy can directly output the probability of an action instead of the mean and co-variance of our action distribution. The policy therefore changes from π: S $\rightarrow R^{2|A|}$ to π: S $\rightarrow [0,1]^{|A|}$ where now we are applying a soft-max function in the final layer of the policy to ensure it outputs a valid probability distribution;

- The soft Q-function cost $J_Q(\theta)$ is minimized by calculating the expectation directly instead of plugging in our sampled actions from the replay buffer to form a Monte-Carlo estimate of the soft state-value function. This change should reduce the variance involved in our estimate of the objective function $J_Q(\theta)$. Then, the state-value function can be expressed as:

$$V(s_t) = \pi(s_t)^T[Q(s_t) - \alpha log(\pi(s_t))] \tag{2.11}$$

- Similarly, we can make the same change to our calculation of the temperature loss to also reduce the variance of the estimate. The temperature objective changes from Eq. 2. to:

$$J(\alpha) = \pi_t(s_t)^T[-\alpha(log(\pi_t(s_t)) + \overline{H})] \tag{2.12}$$

- The reparameterization trick is not needed anymore to minimize $J_\pi(\phi)$, now our policy outputs the exact action distribution and it is possible to calculate

the expectation directly. The policy losses changes to:

$$J_\pi(\phi) = E_{s_t \sim D}[\pi_t(s_t)^T[\alpha log(\pi_\phi(s_t)) - Q_\theta(s_t)]] \tag{2.13}$$

The algorithm for SAC with discrete actions is given by Algorithm 1.

Algorithm 1 Soft Actor-Critic with Discrete Actions

Initialise $Q_{\theta_1} : S \rightarrow \Re^{|A|}, Q_{\theta_2} : S \rightarrow \Re^{|A|}, \pi_\phi : S \rightarrow [0,1]^{|A|}$
▷ Initialise local networks
Initialise $\overline{Q}_{\theta_1} : S \rightarrow \Re^{|A|}, Q_{\theta_2} : S \rightarrow \Re^{|A|}, \pi_\phi : S \rightarrow [0,1]^{|A|}$
▷ Initialise target networks
$\overline{\theta}_1 \leftarrow \theta_1, \overline{\theta}_2 \leftarrow \theta_2$ ▷ Equalise target and local network weights
$D \leftarrow \emptyset$ ▷ Initialise an empty replay buffer
for each iteration **do**
  for each environment step **do**
    $a_t \sim \pi_\phi(a_t|s_t)$ ▷ Sample action from the policy
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$ ▷ Store the transition from the environment
    $D \leftarrow D \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$ ▷ Store the transition in the replay buffer
  end for
  for each gradient step **do**
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J(\theta_i)$ for $i \in \{1,2\}$ ▷ Update the Q-function parameters
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J(\phi)$ ▷ Update policy weights
    $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$ ▷ Update temperature
    $\overline{Q}_i \leftarrow \tau Q_i + (1-\tau)\overline{Q}_i$ for $i \in \{1,2\}$ ▷ Update target network weights
  end for
end for
return $\theta_1, \theta_2.\phi$ ▷ Optimized parameters

Chapter 3

Case study and control problem

In this chapter, a brief description of the control problem is provided. A multi-energy office building with electrical and thermal load is the object of the control policy which must define the operations scheduling of a cold water storage tank and a Li-ion Battery Energy Storage System.

A. Amato et al.[85] already drawn the 3D model of the facility using the OpenStudio plug-in for the design software SketchUp. Then, the SketchUp output file was converted into an Input Data File (IDF) file, editable by the building energy simulation program EnergyPlus. Using EnergyPlus, the building model was completed with the definition of its envelope and all the assumptions relevant for assessing the future energy performance of the facility in operation, also built using eco-sustainable and innovative materials.

The thermal zones are served by an HVAC system that delivers cooled air by means of fun coils and can be fed by an electric chiller or the thermal storage. The HVAC system was simulated using EnergyPlus. The interaction between EnergyPlus and

the controllers in Python was achieved using the Building Control Virtual Test Bed (BCVTB) as a middleware. The electric chiller is implemented in the EnergyPlus environment by defining its nominal operating conditions and the equations that relate the cooling capacity and the COP as functions of the leaving chilled water temperature, the entering condenser fluid temperature and the Partial Load Ratio. The chiller can provide cooling energy to the building, to cold water storage or to both of them. The tank can be seen as a buffer between chiller and HVAC aiming at decoupling the chiller electricity consumption and the cooling load served to the zones. The circulation pump is active whenever the chiller is on or the tank is discharging.

The cooling system can operate in two modes, i) Charging mode, where the cold water is fed to the storage tank by the chiller, and to the building in case the cooling load is requested, ii) Discharging mode, where the demand of the building is met only through the storage, if needed, otherwise no water is circulated and the circulation pump is turned off. The controller chooses among these two configurations. The Fig. 3.1 represents the two operational modes, during the charging phase, a three-way valve regulates the fraction of flowrate that bypasses the tank according to the building cooling load, the remaining is sent to the tank, and it is given by the difference between the chiller cooling capacity and the cooling load:
In discharging mode the chiller is by-passed and the building is cooled only via the cold thermal storage. Both have a fixed flow rate. Moreover, safety constrains are introduced in order to guarantee that the cooling demand of the building is always met and to maintain the temperature of the storage within the prescribed range, also the cooling capacity of the chiller limits the amount of energy transferable to the storage and, in discharging mode if the temperature of the storage tank rises above the maximum and the building cooling demand is not zero the system automatically switches to charging/chiller cooling mode in order to meet

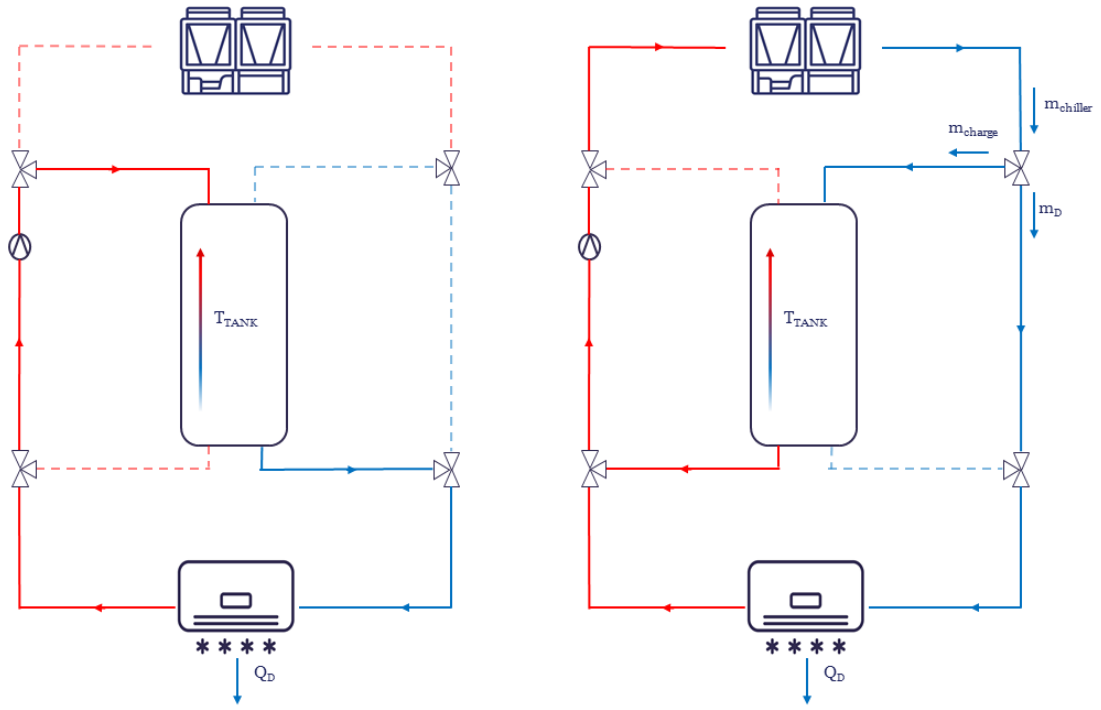

Figure 3.1: Cooling loop: Discharging mode on the left-hand side, charging mode on the right-hand side. The blue lines represent the supply lines, the red lines represent the return lines. Dashed lines are on idle

the demand regardless of the control action foreseen by the agent. For simplicity, the building's thermostatic control is not considered and its cooling demand is considered as an external disturbance along with the price of the electricity and the zone temperature where the storage is located.

The electrical load of the building is composed by the electric chiller and the circulation pump and it is treated as aggregated. The test facility will be used by students and staff in the central hours of the day, akin to a typical tertiary sector building. Thus, a good matching between energy consumption and PV generation power profiles is expected. For this reason, PV generators will be installed to supply the electricity demand. The PV model is implemented into a Python class and solar positions are imported from the pvlib library. The operation of BESS was

also simulated with an energy model developed in the Python environment. The most used battery models involve the estimation of the State Of Charge (SOC), which is a simple and easily implemented method, but it is sufficient for carrying out a preliminary evaluation on the impact of BESS installation, even though the degradation of the battery is not taken into account. The implemented model is based on the calculation of the SOC at any time instant t according to the set of equations in Eq. 3.1:

$$\begin{cases} SOC(t) = SOC(t-1) - \eta_{rte}\frac{P_{batt}(t)\Delta t}{C_{batt}} & if P_{batt} < 0 \\ \\ SOC(t) = SOC(t-1) - \frac{P_{batt}(t)\Delta t}{C_{batt}} & if P_{batt} > 0 \end{cases} \tag{3.1}$$

where SOC$(t-1)$ is the SOC at the previous time instant $t-1$, η_{rte} is the round-trip efficiency , P_{batt} is the average power exchange in the time-step Δt between the battery and the system ($P_{batt} < 0$ in charge and $P_{batt} > 0$ in discharge) and C_{batt} is the battery nominal energy capacity. Moreover, the battery has safety constraints in order to preserve its lifetime. Charging and discharging processes have to respect two limits $|P_{charge}| < P_{charge,max}$ and $|P_{discharge}| < P_{discharge,max}$, defined in the technical specifications and required to avoid too fast charging/discharging. Typically, maximum charging and discharging power are different and when the power exceeds these thresholds, the controller limits it to the maximum recommended value P_{max}. Limits on the State Of Charge must be respected and the battery SOC must not exceed the minimum and maximum values provided by the manufacturer, i.e. $SOC_{min} \leq$ SOC $\leq SOC_{max}$, in order to preserve the health of the battery. The energy surplus or deficit lead to energy exchanges between the facility and the external electrical grid (energy injections and absorptions, respectively).

The electrical system is composed by a DC bus and an AC bus, interfaced by a mono directional DC/AC inverter. The PV is connected to the DC bus by a DC/DC

converter and can inject electricity to the building, to the battery and to the grid, ordered as the descending level of priority. BESS exchanges electricity with the DC bus through a DC/DC converter and the ratio between energy discharged and energy charged is equal to the round-trip efficiency net of the converter efficiency, given that there are no internal losses. Grid is not allowed to charge the battery according to the normative of many European countries, but it is used to assist in matching demand and generation at each instant[86], so it is assumed that the grid is always able to balance the electrical system. The Fig. 3.2 represents the electrical system, along with allowed directions for each flow:

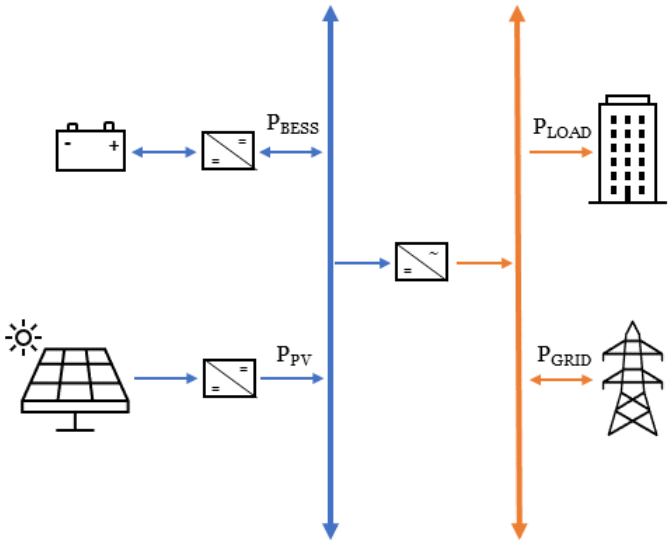

Figure 3.2: Electrical layout

At each time-step, the electricity exchanged with the grid can be positive or negative, so the energy cost is computed according to a buying price schedule and, a selling price schedule for electricity injected to the grid. The schedule is based on the Time of Use fare plan with low, medium and peak price hours, and it is known in advance for the whole season. Feed-in tariffs are out of the scope of this work, thus the selling price is fixed over the whole control period.

The goal of the control policy is to minimize the energy cost by exploiting the energy storage to increase the building flexibility, and perform peak shifting to low price hours. The control policy is realized by defining the operation of the energy storage given a set of information. This work makes use of two very different control strategies, which are referred to as the baseline and the proposed strategy. The baseline adopts two Rule-Based Control strategies to manage both the energy storage, whereas the proposed strategy uses a smart agent from the SAC algorithm to control the thermal storage and has as the same RBC as the BESS. In the following section, these control strategies are described.

3.1 Rule-Based Control

The thermal storage follows a peak shifting strategy where it acts as a buffer between the chiller operation and cooling load. During low price hours, the chiller charges the tank at the maximum flow rate according to pre-defined temperature set-points, which are set to 10°C and 12°C. When the price is the highest the tank starts discharging until the storage temperature reaches the maximum temperature or the building cooling demand is null. During the week-end, the building load is zero, but the chiller keeps charging the tank in order to take advantage of the PV production and to discharge the battery.

A classical RBC from the literature is used to manage the BESS, which is used by the vast majority of the authors. This strategy is considered as the most simple, but still very effective. Both Ruusu et al.[87] and A. Amato et al.[85] suggest this straightforward approach, where battery is charged when PV production excesses the building load otherwise it is discharged. Being more specific, during charging the PV surplus is diverted to the battery if it is allowed by the constraints on charging power and maximum SoC, otherwise the grid draws the remaining energy fraction. During battery discharging, the battery works in parallel to the PV to

meet the demand, if the contribution from both PV and battery is not enough, the grid guarantees that the building load is satisfied. This strategy is implemented in both the baseline and the proposed strategy. In Fig. 3.3, the flowchart representing the battery RBC is shown.

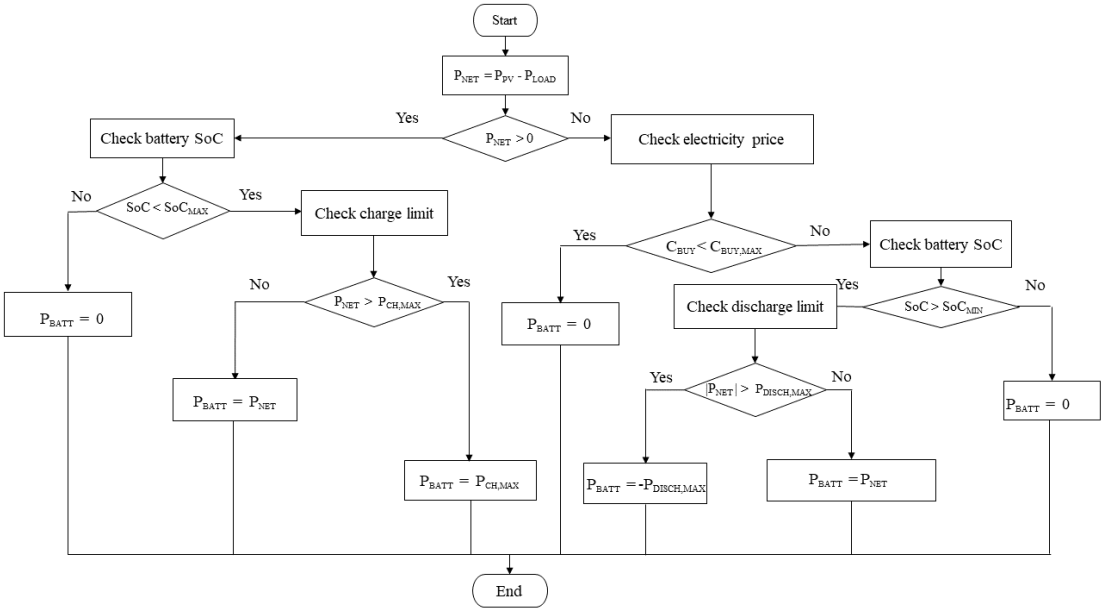

Figure 3.3: Rule-Based Control strategy for BESS

3.2 Reinforcement Learning control

The proposed strategy exploits an advanced control strategy for managing storage tank operation. It aims at scheduling the chiller by choosing between the charging and discharging phase of the thermal storage. At each control time-step, the control agent is fed with the observed state from the environment, processes it and returns back the chosen action to the environment. Eventually, the control action is adjusted by the environment according to a few constraints in order to make sure the thermal load is satisfied, the upper and lower temperature limits are not

violated and the maximum chiller capacity is not exceeded. Finally, the control action is performed in the simulation and repeated over until the next control time-step. Again the charge and discharge flow rate are fixed.

Chapter 4

Implementation

4.1 Case study

The test facility analyzed in this work was devised within the PhotoVoltaic Zero Energy Network (PVZEN) research project at the Politecnico di Torino, thanks to the co-operation o fthe Department of Energy (DENERG), the Department of Architecture and Design (DAD) and the Department of Electronics and Telecommunications (DET). The goal of the project is to build an all-electric nZEB that fulfils its energy demand through PV systems and uses BESS to be independent from the external electrical grid [85]. It consists of two study rooms, one control room and a technical room. The technical room is not served by the air-conditioning system and the storage tank is placed within it. Rooms are separated by partition walls. It is a prefabricated building with a rectangular layout. The area occupied by the facility is 196.3 m^2 (11.25 $\times$ 17.45 m), including an outdoor zone with wheelchair ramp and steps and a shed-covered zone at the glazed entrance doors of the study rooms and the control room. the interior air-conditioned area is around 96.8 m^2. The ceiling height is 2.8 m at the minimum and 3.7 m at the maximum above the floor level, due to the different tilt angles of the roof, which are 13.4° on SE

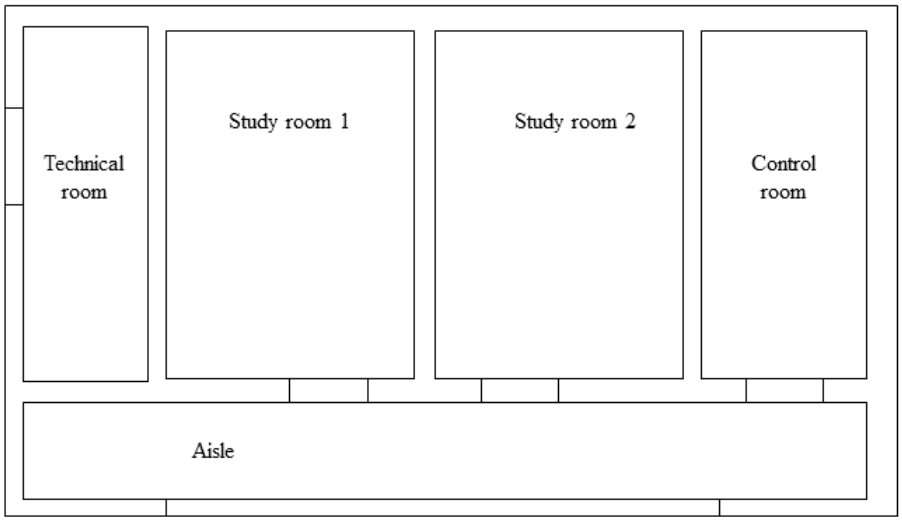

Figure 4.1: Building layout

side and 15° on NW side. In Fig. 4.1, the facility layout is presented, showing the distribution of the rooms. The equipment and the size of the rooms are different as they have been designed for carrying out different kind of activities. The useful area and the function of each room are described in Tab. 4.1 and the building envelope is characterized by the construction features listed in Tab. 4.2. The weather file used in this work is the reference weather file (ITA TORINO-CASELLE IGDG.epw) available in EnergyPlus for Torino, Italy which provides outdoor temperature and solar irradiance collected by the weather station at Caselle Airport. The input data represents the cooling season which is defined from June to August. The rooms

Rooms	Area	Use
Technical room	15.7 m^2	Location of electronic and HVAC equipment
Study room 1	30.4 m^2	Hosting students
Study room 2	30.4 m^2	Hosting students
Control room	20.3 m^2	Monitoring of energy systems and performance

Table 4.1: Test facility rooms

are served by an HVAC system to satisfy the cooling load, which can be fed by an

Feature	Value
Conditioned floor area	96.8 m^2
Conditioned volume	501 m^3
Envelope surface/conditioned volume ratio	0.85 m^1
Transparent/opaque envelope surface ratio	6.6%
Opaque envelope surface	400 m^2
$\hat{U}_{op}$	0.16 W/m^2K
$\hat{U}_{tr}$	0.55 W/m^2K

Table 4.2: Bulding envelope characteristics

electric chiller or a cold-water storage tank. The cooling demand was considered as an external disturbance of the system and was calculated within EnergyPlus in order to maintain an indoor temperature of 26 °C and a relative humidity of 55% between 08:30 and 18:00 from Monday to Friday and the air ventilation rate for the control room and the study rooms was based on the occupancy at 10 L/s per person, thus resulting in 30 L/s and 100 L/s, respectively.

The chiller is chosen with a reference capacity Q_{cap} of 12 kW and the reference COP of 2.67. The design water mass flow rate during charging phase is 0.2 kg/s while during discharging phase is 0.35 kg/s. This latter value correspond to the sum of the design mass flow rates of the three air-conditioned zones. The supply water temperature at the outlet of the chiller was set equal to 7 ℃. The tank operates in the range between 10°C and 18°C which correspond to a State Of Charge of 1 and 0, respectively. The thermal losses are computed from the global heat transfer coefficient and the temperature difference between the tank and the technical room air. The value for this parameter is obtained from technical datasheets.

A commercial monocrystalline silicon photovoltaic module will be mounted, namely BP Solar BP 585 F. The selected module has a specific power of about 140 W/m^2 and an efficiency of 15%, under standard test conditions (solar irradiance G_{STC} = 1000 W/m^2 , cell temperature T_{STC} = 25°C, $AirMass_{STC}$ = 1.5), as described by

Durisch et al.[88] and reported in Eq. 4.1.

$$\eta = f(G, AM, T_{out}) \tag{4.1}$$

The PV panels tilt angle has been chosen from the world dataset provided by M.Z. Jacobson and V. Jadhav[89] which perform the optimization for a large amount of location. Among those location, Lyon is selected as the most suited for this case study and the tilt is set to be as high as 33°, whereas the azimuth is constrained by the orientation of the test facility. These inputs along with solar radiation and incidence angle allows to compute the PV specific power at each simulation time-step. The PV field design takes into account the total demand from the cooling season only, in this way the nominal power of the modules is chosen in order to match up to the peak power of the building total demand. The Tab. 4.3 recaps the parameter of the PV module.

Parameter	Value
Nominal power	3 kW
Surface	22 m^2
η_{STC}	0.15
Tilt angle	33°
Azimuth angle	116°

Table 4.3: PV parameters

The characteristics of the BESS considered in this work were provided by the datasheet of a modular Li-ion battery available on the market and are shown in Tab. 4.4. Several nominal capacities are explored in the simulation. A SOC_{min} = 10% and a SOC_{max} = 80% were assumed for a total Depth of Charge of 80%, in compliance with the typical values for the lithium-ion technology. An initial SoC of 50% was set. Maximum charging and discharging power are set to 0.5 C and 1 C. During the opening hours, the zones were supposed to be occupied at their maximum capacity, which means the control room and the two study rooms were

Parameter	Value
Round-Trip Efficiency	0.96
Maximum discharging power	1C
Maximum charging power	0.5C
SOC min	10%
SOC max	90%

Table 4.4: BESS characteristics

assumed to be constantly occupied by 3 and 10 people, respectively. No regular occupation was expected for the technical room. The air infiltration rate was set to 0.15 vol/h, a typical value for office buildings. The control time step and the simulation step are set equal to 1 hour. Obviously, this choice reduces the accuracy of the simulation, but the computational cost is dramatically lower. The electricity cost is scheduled for buying and selling: across Europe there are many examples of TimeOfUse tariffs, which have been successfully implemented, and it is at the moment the most spread DR program available for small users. The price of the electric energy drawn from the grid to operate the chiller unit is based on the tariff structure commonly implemented in Italy. The simulation time is divided into low price, medium price and high price periods. Specifically the low and medium price values were chosen to be 1/10 and 1/2 of the higher one, respectively, corresponding to 0.03 €/kWh, 0.165 €/kWh and 0.3 €/kWh. On week-days the price is low from 0:00 A.M. to 7:00 A.M. and from 11:00 P.M. until 0:00 A.M., the medium price period goes from 7:00 A.M to 8:00 A.M. and from 6:00 P.M. to 11:00 P.M. and finally, the high price period goes from 8:00 A.M to 7:00 P.M.. On saturday, the price is medium from 7:00 A.M. to 11:00 P.M. and it is low before 7:00 A.M. and after 11:00 P.M.. On sunday, the price is always at low fare. The price of the electricity were assumed relatively to the maximum price, in order to discriminate the values for the optimization application. The Fig. 4.2 visualizes the fare scheme.

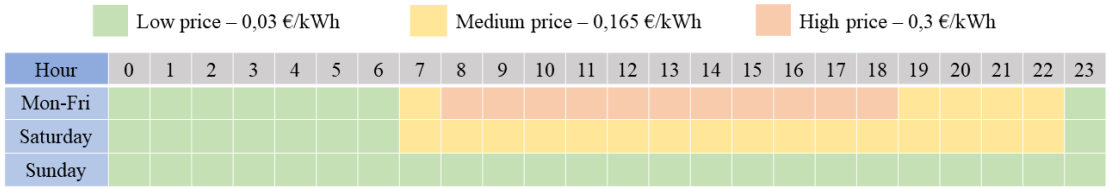

Figure 4.2: Electricity TimeOfUse tariff

Regarding the selling option, the italian GSE refers to "scambio sul posto" to compute the revenue for production from renewables; in case of PVs, the GSE contributes with its "Contributo in conto scambio" on a yearly basis, in this work, since there are no data for the whole year, it is assumed it to be equal to 0.01 €/kWh[90]. The efficiency of monodirectional DC/AC is assumed to be equal to 90% and the efficiency of DC/DC converters to 95%.

4.2 Design of Reinforcement Learning

One of the most recently researched Reinforcement Learning algorithm for energy systems is the Soft-Actor-Critic with replay buffer and target network. These two last improvements have raised the possibility for actor-critic to play a role in the building control field. The reinforcement learning control is used in this work to solve a highly non-linear problem, and it is designed by defining the action space, the reward function and the state space. Besides the formulation of the reward function and of the state-space, the reinforcement learning frameworks requires of a series of hyperparameters to be set as the discount factor for future rewards γ and the structure of the neural networks employed as function approximators. The length of the learning phase is very important to assess the applicability of RL control and is influenced by these values and by the definition of the observation space and the reward function as well. The values of the hyperparameters selected for this application are summarized in Tab. 4.5.

Hyperparameters	
Discount factor	0.99
Actor network learning rate	0.0005
Critic network learning rate	0.0005
Soft-update Boltzmann Temperature coefficient	0.005
# hidden layers	3
# neurons in hidden layer	512
batch size	64
Reward scaling	10
Reward weight factor	100

Table 4.5: Hyperparameters for SAC training

This algorithm will map the optimal policy to follow given a certain state, which is a discrete probability distribution where actions are sampled from. The training goes on by repeating the same training episode in order to let the agent converge to the optimal control policy. Once the agent is trained, the deployment phase takes place to assess the performance of the policy. The deployment can be static or dynamic, which defines whether the networks stop being updated or not while the agent keeps sampling action from the Gaussian distribution. For the purpose of this work, the condition of the environment are kept as the same as in the training phase, so that the agent does not have the need to keep updating the weights and biases of the network. In the following paragraphs, the definition of observation space, action space and reward function are presented.

4.2.1 Observation space

The agent choose the action given a set of information provided by the observation space, which has to be as representative as possible of the current state of the system. Normalization is needed to process output from the simulation before feeding to the algorithm.

In this study the state space does not include only information about of the current time step but also information about the recent past and future disturbances. The storage tank State Of Charge (SOC) at the timestep k was introduced to provide to the agent information about the amount of energy actually stored. Moreover, past values of this variable were introduced to provide information about the evolution of the temperature caused by the charging/discharging of the system up to 2 hours before the current control time step. The building cooling demand together with the price of electricity is a fundamental information to optimally manage the controlled system. The electricity price is a key-information for the agent in order to recognize peak price hours and correctly plan the operations of the system. Present value is provided along with the exact values for the 24 hours ahead. The electricity price patterns were supposed to be always known, thus no dynamic pricing is taken into consideration. Also, the values of building cooling demand from time step k to time step k + 24 were provided to the agent. The predictions of building cooling demand were assumed to be deterministic, but future works could use neural networks model them, since many applications in energy forecasting have been studied recently. The outdoor air temperature affects the COP of the chiller unit so it is included as well. Moreover, the agent must know the amount of on-site generation available, so that PV production and its predictions for the next 24 hours and the State Of Charge of BESS are fed as input. This gives the agent the ability to exploit the chiller during high irradiance periods or when battery is fully charged.

The Tab. 4.6 shows the state variables along with their maximum and minimum value.

All the variables included in the state-space are physical quantities directly extracted from the simulation output with the exception of the State of Charge (SOC) of the storage tank that was calculated according to Eq. 4.2:

State variable	Min	Max	Unit
Outdoor temperature	7	40	°C
Tank SOC	0	1	
Tank SOC 1h lag	0	1	
Tank SOC 2h lag	0	1	
Battery SOC	0	1	
Cooling load	0	10000	kW
Cooling load 24h prediction	0	10000	kW
PV generation	0	2000	kW
PV power 24h prediction	0	2000	kW
Electricity price	0.03	0.3	€/kWh
Electricity price 24h prediction	0.03	0.3	€/kWh

Table 4.6: Observation space

$$SOC(t) = 1 - \frac{T_s(t) - T_{s,min}}{T_{s,max} - T_{s,min}} \tag{4.2}$$

4.2.2 Action space

At each control time step, the agent has to define the scheduling of the charging and discharging modes. SAC algorithm can work with continuous action space, but it involves the exploration of a very wide range of action-space pairs before converging, hence the learning could become overly large and the applicability in real-time operation is discouraged. Given the motivation for shorter learning time, the action space is selected to be discontinuous. Particularly, the controller must choose between 1 and -1, which correspond to charging mode and discharging mode of the cold-water storage tank, respectively. This implementation reduces the state-action pairs to be explored, without major drawbacks. The control action must respect the safety constraints otherwise the system operation is adjusted in order to meet them.

4.2.3 Reward function

The reward function obtained by the agent after selecting an action at each time step measures its control performance. The objective of the controller is to minimize the energy cost, which is positive when energy is withdrawn from the grid, but the use of PV allows negative costs when its production excesses the aggregated building electricity demand and battery is already fully-charged. The reward function is expressed by the Eq. 4.3:

$$\begin{cases} r(t) = \beta E_{grid}(t) \cdot C_{buy}(t) & if E_{grid}(t) < 0 \\ r(t) = \beta E_{grid}(t) \cdot C_{sell}(t) & if E_{grid}(t) > 0 \end{cases} \tag{4.3}$$

Where E_{grid}(t) refers to the electricity exchange between the facility and the external grid at time-step t, C_{buy}(t) and C_{sell} are defined according to the schedule price for buying and selling electricity and β is a weight factor introduced to regulate the magnitude of the reward, namely reward scale, and it is considered an hyperparameter of the algorithm.

4.3 Simulation environment

The system model and the control strategy are implemented in a simulation environment which enables the information exchange between EnergyPlus v9.2.0 and Python v3.0. The Building Control Virtual Test Bed (BCVTB) and the ExternalInterface-Ptolemy server command from EnergyPlus were used to connect the two software. EnergyPlus is required, to perform the simulation of the energy system, while Python is implemented with the control agent, the latter based on OpenAI Gym. For the control phase, an external interface had to be used to

implement the algorithm within the EnergyPlus simulation. The environment has 4 main functions in Python:

- init(), where the Python class is defined;
- step(), which receives a specific action, and then it implemented in the simulated building to return a tuple composed by the next state, reward, done (True/False) and info;
- reset(), a function that is called at the beginning of each episode, to start over by returning the initial state;

An episode corresponds to a whole simulation of the cooling season and it is needed to be repeated several times during the training phase to get acceptable results. The information exchange between the two software is as shown in the Fig. 4.3 The process can be resumed as follow: the OpenAI gym interface object is initiated by calling the init() function, then a server socket for the communication between EnergyPlus and Python is created; the reset() function is called up by the control agent immediately afterwards and an instance of EnergyPlus is immediately created using the IDF file format and the CFG extension file that allows data exchange. The OpenAI Gym object creates a TCP connection with EnergyPlus, in which ExternalInterface incorporates features that are inputs from Python. The ExternalInterface uses a BCVTB to perform as a client. The TCP connection is used to read and return the simulation output from EnergyPlus to OpenAI Gym. Then observations are processed by DRL agent for extracting state and reward. The DRL agent calls the step(a) function at each control steps and sends the action to Energyplus for each simulation step and reads the results. Generally, the control time-step is larger than the simulation time-step, but in this work they have the same length. Finally, the observations are returned in order to obtain new state and reward. At each time-step, the OpenAI checks if the simulation is at the end

of the episode: if it happens, the process moves on to the next check, otherwise, it is repeated the above process starting from the observations obtained again by EnergyPlus; if the processed episode is the last one, then the process ends here. Otherwise, it starts again from the point where the reset() function is called.

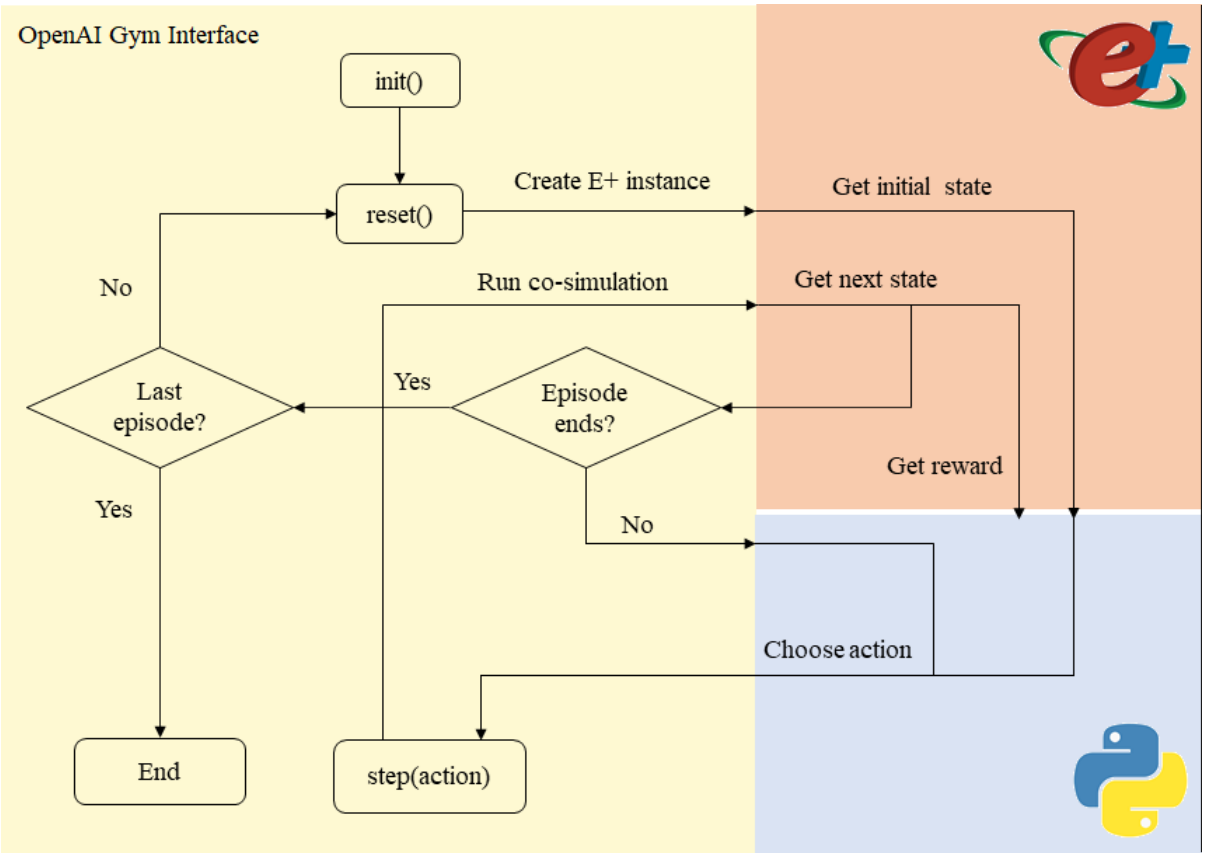

Figure 4.3: Simulation environment

4.4 Experimental setup

The simulation is carried out for several configurations, in order to analyze the impact of the different designs of the energy system. The aim is to find out how the proposed control policy can achieve the same or better results with respect to

a classical control while utilizing lower size equipment and thus saving economic resources which is decisive to guarantee the spread of the storage technologies.
The variables to be investigated are the volume of the cold-water storage tank and the nominal capacity of the BESS.
These configurations are implemented in the environment and the SAC algorithm is trained for all of them while keeping the other variable as constant. In the deployment phase, the trained algorithm performs the control statically and the performance can be assessed for the discussion of the result. The thermal capacity of the tank is a function of volume and temperature, in this case, the temperature operation range is kept constant while the volume is subjected to changes for three different scenarios. Also, larger tanks involve higher exchange surface, so that the thermal loss increases. Eventually, the heat gain coefficient is scaled for different surface. The Tab. 4.7 reports the values of the tank size implemented in the environment.

Volume	Heat gain
10 m^3	12 W/K
8 m^3	10.3 W/K
6 m^3	8.5 W/K

Table 4.7: Cold-water tank design

The use of electrical storage is not encouraged from an economic point of view. High investment cost and low profitability characterize the current market[91–94], but costs are expected to drop more and more in the next year and also subsidies for purchasing electrical storage technologies will push their penetration. Nevertheless, BESS improves PV energy performance, by addressing the problem of the solar energy low flexibility. Nominal capacity is explored for 2.4 kWh, 4.8 kWh and 7.2 kWh.
The configurations are named and resumed in Tab. 4.8.

Name	BESS capacity [Wh]	TES volume [m^3]
1	2400	10
2	4800	10
3	7200	10
4	2400	8
5	4800	8
6	7200	8
7	2400	6
8	4800	6
9	7200	6

Table 4.8: Configurations simulated for the experiment

The same episode is repeated over for 30 times at each configuration, during which, the networks weights and biases are updated. Eventually, the parameters from the best episode are saved and used in the deployment phase. Since the environment is as the same as in the training phase, the deployment is static. The next chapter shows and discusses the results of the simulation.

Chapter 5

Results and discussions

The results from the simulation environment are resumed by several graphs and tables in order to depict the the different scenarios and to find out what are the main implication of the these control strategies.

5.1 Results

First, the global performance of the two strategies is assessed, which is related to the operating cost. For case study, it is computed as the seasonal expenditure for exchanging electricity with the distribution grid. Energy cost ranges from 4.7 € to 10.1 € when using RLC and from 12.4 € to 18.1 when using RBC. RBC operational cost decreases as the BESS size increases, as expected. It achieves a reduction between 26.1% and 31.5% when nominal capacity is pushed up to 7200 Wh from 2400 Wh. RL has not such behaviour due to the intrinsic randomness, but still it achieves a reduction between 14.0% and 44.7% from 2400 Wh to 7200 Wh.
Obviously, the best performance occurs when the BESS has the highest nominal capacity for both the control strategies. Specifically, RB cost is not influenced if the BESS is very large, so that the worst configuration has a 2400 Wh BESS and

a 6 m^3 TES, while the best has a 7200 Wh BESS and a 6 m^3 with a total cost of 12.4 €. RL has still an anomalous behaviour, with the configuration 6 as the optimal one with a cost of 4.7 €. These results are resumed in Fig. 5.1.
RLC always performs better than RBC as shown by Fig. 5.3, in terms of energy cost and energy consumption. Particularly, RLC allows between 31.4% and 62.5% cost savings with respect to RBC. The energy consumption shows a little difference, since it is strictly related to the cooling load which depends on the fixed external conditions. Nevetheless, RLC is able to reduce the energy consumption, even though the energy savings are less significant than the cost savings. Energy consumption is shown in Fig. 5.2. RBC energy consumption only varies with the tank size, and it decreases from 1090.7 kWh to 1072.2 kWh when the volume is reduced from 10 m^3 to 6 m^3, with a 1.7% drop. On the other hand, the RLC energy consumption seems more stable across all of the configurations, where values ranges from 1064.9 kWh and 1078.1 kWh. As a consequence, the energy savings from RLC are higher at 10 m^3 storage tank, but still their value ranges from 1.2% and 1.6%. Configuration 6 turns has not the least energy consumption, even though it has the least cost.

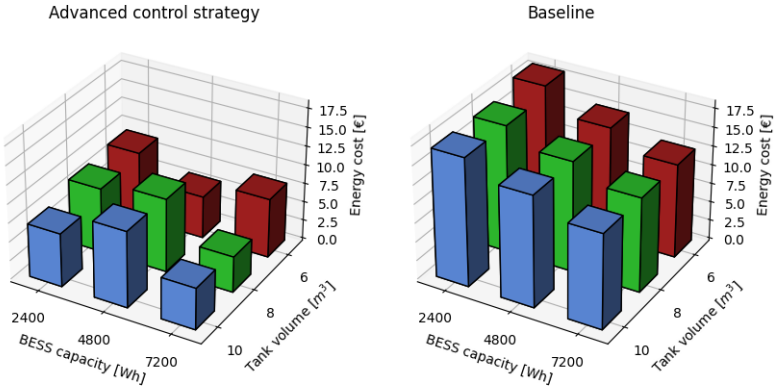

Figure 5.1: Energy cost across all configurations for RB and RL

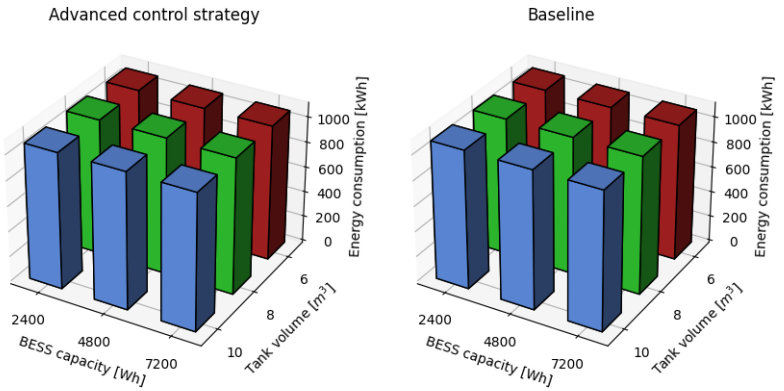

Figure 5.2: Energy consumption across all configurations for RB and RL

Each simulation can discussed in terms of grid interaction. As follows, Tab. 5.1

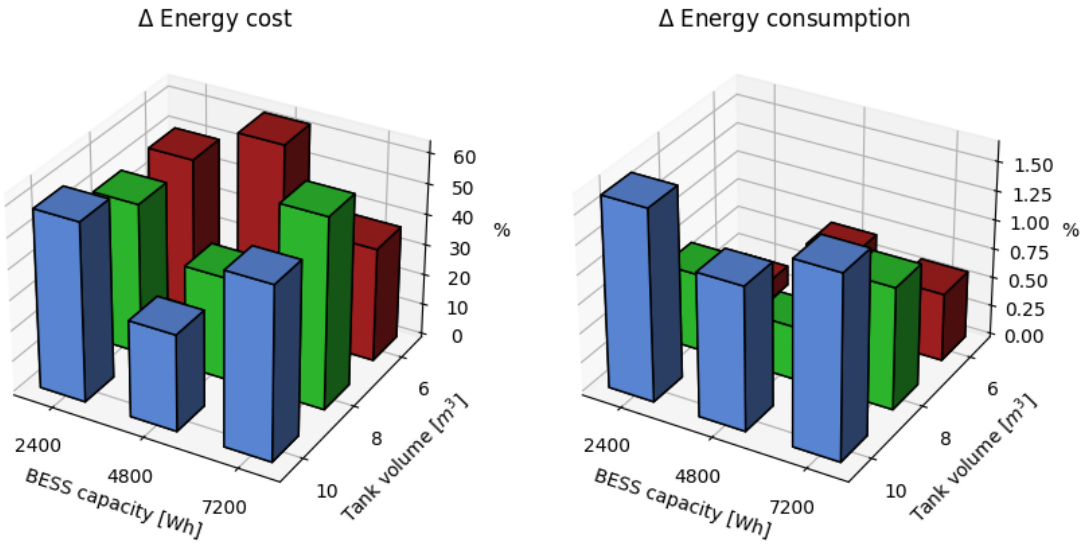

Figure 5.3: Percentage difference between RL and RB across all configurations for energy consumption and energy cost

and 5.2 show the results of the two different control strategies Both strategies

ID	En. bought	En. sold	Buying cost	Selling cost
	kWh	kWh	€	€
1	871.4	919.1	26.1	9.2
2	749.1	776.8	22.5	7.8
3	628.5	636.5	18.9	6.4
4	872.7	928.1	26.2	9.3
5	750.9	786.4	22.5	7.9
6	632.0	648.0	19.0	6.5
7	861.1	928.9	27.4	9.3
8	747.2	796.2	22.9	8.0
9	636.6	667.3	19.1	6.7

Table 5.1: RB results from grid operation for all configurations

reduce either the energy bought and the energy sold when BESS size is increased, Especially, the energy requested to the grid sees an average reduction of 27.1% and 45.1% from 2400 Wh to 7200 Wh for RB and RL, respectively. This is enforced by the ability of RL of reducing by 42.7 % on average the energy sold to grid, against 29.7% for RBC. RBC always buys electricity during low price hours except in configuration 7 where at the end of the opening time the system is forced to withdraw energy from the grid to satisfy the cooling load, due to the small TES volume. The specific purchasing cost of electricity is 26.2% on average higher in the case of RLC, but this is compensated by reducing for 66.0% on average the amount of energy drawn from grid and resulting in a reduction from 42.3% to 64.1% of the buying cost.

The energy balance of the TES is resumed by the Tab. 5.3 and 5.4. The BESS capacity is not influencing the charge/discharge of the storage tank, in the case of RBC. As the tank size increases, the energy discharged reaches a maximum where the cooling load is always met by the tank, while the thermal losses keep increasing. From 8 m^3 to 10 m^3, the energy discharged increases by 0.1% with thermal losses being 13.3% higher. Differently, both BESS and TES size influences

ID	En. bought	En. sold	Buying cost	Selling cost
	kWh	kWh	€	€
1	333.7	398.5	11.1	4.0
2	236.3	288.0	13.0	2.9
3	203.3	249.1	8.0	2.5
4	353.5	415.2	12.7	4.2
5	278.1	325.3	13.0	3.3
6	162.3	206.4	6.8	2.1
7	335.2	400.2	13.1	4.0
8	232.8	286.0	8.4	2.9
9	193.7	238.6	10.2	2.4

Table 5.2: RL results from grid operation for all configurations

the tank operation when using RLC; indeed, the BESS helps reducing the need of the storage during high price period. The results confirm that the storage operates a lot less thanks to RL, hence the thermal losses are reduced by 11.3% on average, but mainly, energy charged and discharged are reduced by 49.0% and 50.7%. The

Volume	En. charged	En. discharged	Thermal losses
m^3	kWh	kWh	kWh
10	3307.4	3132.8	177.6
8	3281.0	3129.0	153.9
6	3160.2	3046.3	126.5

Table 5.3: Storage tank thermal flows for RBC

management of the electrical flows is better explained by Fig. 5.4, Fig. 5.5 and Fig. 5.6., where the contribution to the energy supply from grid, battery and PV is shown. The changes for different TES volumes are not relevant, so that it is easier to average the results. RBC performs poorly at feeding the building with PV, with figures between 7.8% and 8.8% of the total supply, which are much lower than those in case of RLC, ranging from 54.0% to 58.9%, but mainly, it is implied a shift from grid to battery of the contribution to the energy supply by sizing up the BESS. RLC and RBC shows similar values of BESS supply percentage, which are

ID	En. charged	En. discharged	Thermal losses
	kWh	kWh	kWh
1	1855.2	1726.4	150.2
2	1527.1	1384.7	152.2
3	1629.1	1503.1	140.1
4	1823.2	1686.9	141.6
5	1724.8	1581.7	144.2
6	1656.2	1537.1	128.1
7	1745.0	1621.5	125.8
8	1570.7	1457.6	113.2
9	1401.3	1285.2	115.9

Table 5.4: Storage tank thermal flows for RLC

11.4%, 22.4% and 33.2%, and 13.5%, 21.1% and 27.0%, respectively. By contrast, the average grid contribution goes down from 80.2% to 58.4% and from 31.7% to 17.4% for RBC and RLC, when battery size is increased from 2400 Wh to 7200 Wh.

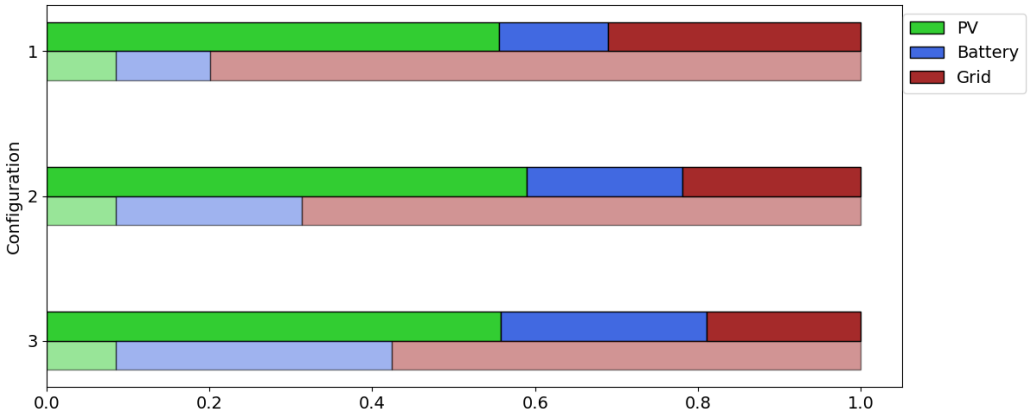

Figure 5.4: Percentage energy contribution from PV, BESS and grid for 2400, 4800 and 7200 Wh with 10 m^3 storage tank, dark color refers to RL and light color to RB.

The main indicators to assess PV-BESS system operation are Self-Sufficiency and Self-Consumption, the former expresses how much of the demand is satisfied by

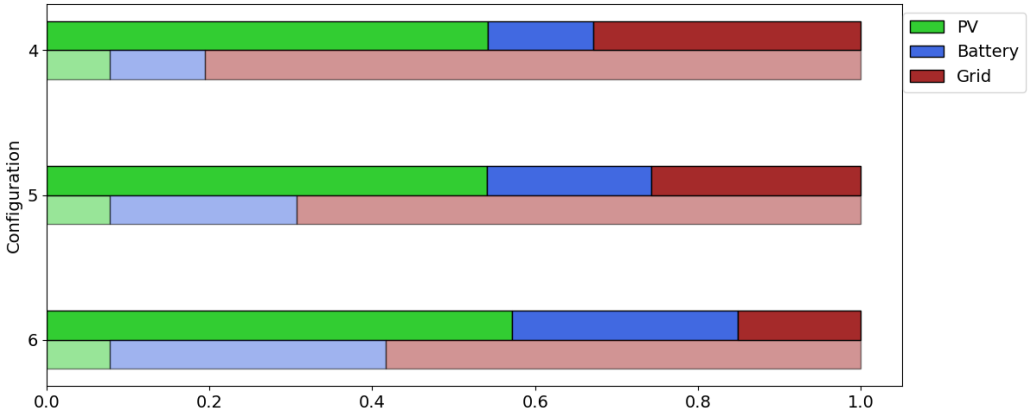

Figure 5.5: Percentage energy contribution from PV, BESS and grid for 2400, 4800 and 7200 Wh with 8 m^3 storage tank, dark color refers to RL and light color to RB.

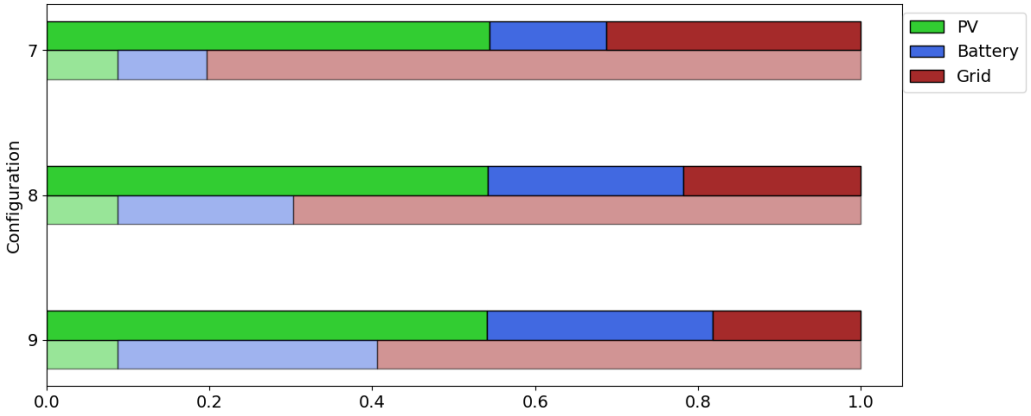

Figure 5.6: Percentage energy contribution from PV, BESS and grid for 2400, 4800 and 7200 Wh with 6 m^3 storage tank, dark color refers to RL and light color to RB.

the local production and the latter expresses how much of the local production is consumed in place. SC is also a good indicator of the economic viability of the PV module which is requested to increase it as much as possible. In this sense, this is what the battery is meant to do. Fig. 5.7, Fig. 5.8, and Fig. 5.9 show the levels of SS and SC; again, TES volume is not affecting significantly these values, so that

the following values are averaged. RLC remarkably gets over RBC; the use of RLC enhances the SS from 19.7% to 68.2%, from 30.8% to 76.8% and from 41.5% to 82.6% and also, SC is pushed from 16.5% to 56.9%, from 25.8% to 64.0% and from 34.9% to 68.6% for 2400 Wh, 4800 Wh and 7200 Wh, respectively. The baseline must have a battery size three times higher in order to reach appreciable levels of performance, but still being far from those reached by the RLC. The difference between the two strategies narrows down as the battery size is increased, but still, the advantage is considerable at 7200 Wh.

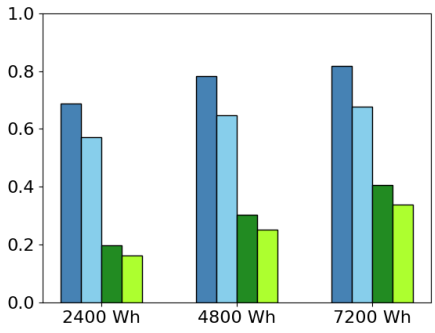

Figure 5.7: SS and SC levels for RL and RB with 6 m^3 storage tank

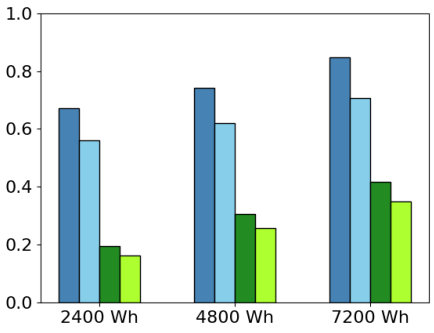

Figure 5.8: SS and SC levels for RL and RB with 8 m^3 storage tank

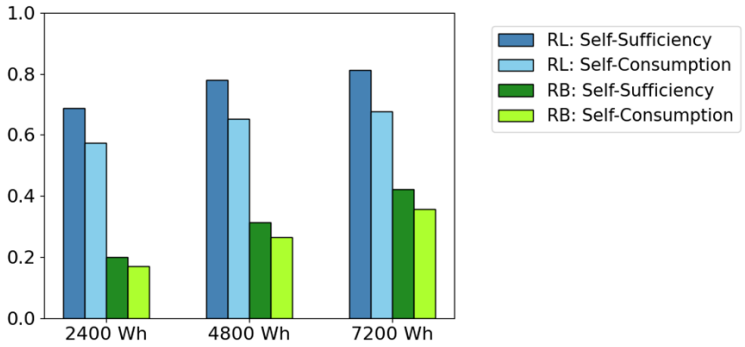

Figure 5.9: SS and SC levels for RL and RB with 10 m^3 storage tank

The two strategies can be investigated more in detail by analyzing them on a daily basis to understand the operational patterns resulting from a Rule-Based

control and a Reinforcement Learning control.

The baseline reports what has been expected, since the rules are already defined and has not any adaptability feature. As mentioned, the chiller charges the tank during the low price hours, in order to use it as a buffer and to efficiently achieve the peak shifting of the building electricity load, hence, RLC does not provide any advantage in these terms. The case study has a low price fare at night so that, as a consequence of the peak shifting strategy, the BESS is discharged at the time the tank is charging, while the PV production exceeds the building load during high price hours charging the battery again. This strategy can be seen in Fig. 5.10 and Fig. 5.11 representing two consecutive week-days. The positive values refer to the energy injected to the building, to the battery or to the grid, whereas the negative ones refer to the energy supplied by the PV module, by the battery or by the grid. The building load is the aggregated energy demand of the chiller and the circulation pump.

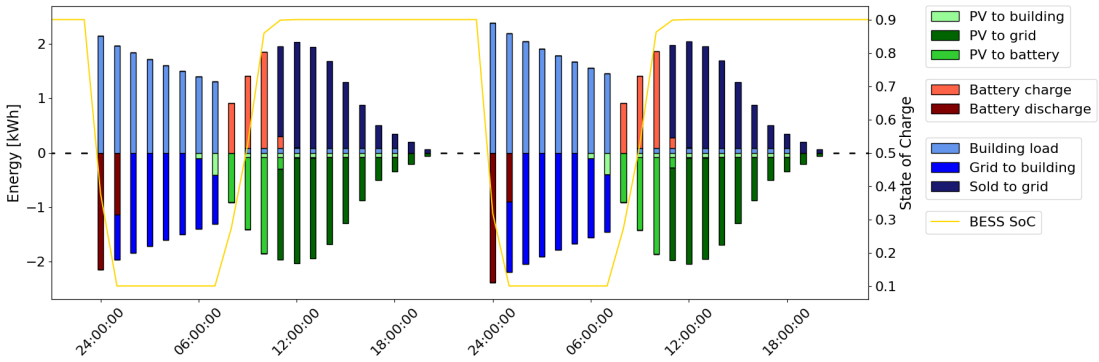

Figure 5.10: Energy flows for all equipment and BESS SoC during two consecutive week-days from 31/06 to 01/07 by adopting RBC at 4800 Wh battery capacity and 8 m^3 storage tank

The system withdraws electricity from the grid only during low price hours, since the storage tank is able to fully satisfies the cooling load during the opening time but this only achieved by making use of a very large tank with respect to the

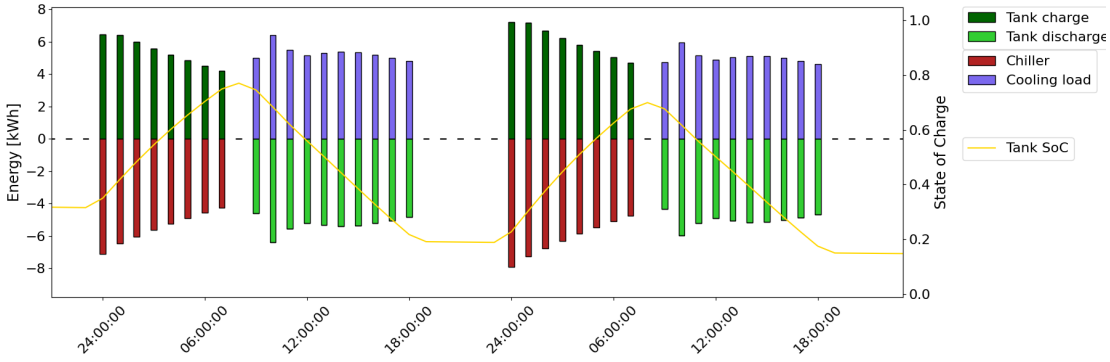

Figure 5.11: Thermal flows for chiller and TES during two consecutive week-days from 31/06 to 01/07 by adopting RBC at 4800 Wh battery capacity and 10 m^3 storage tank

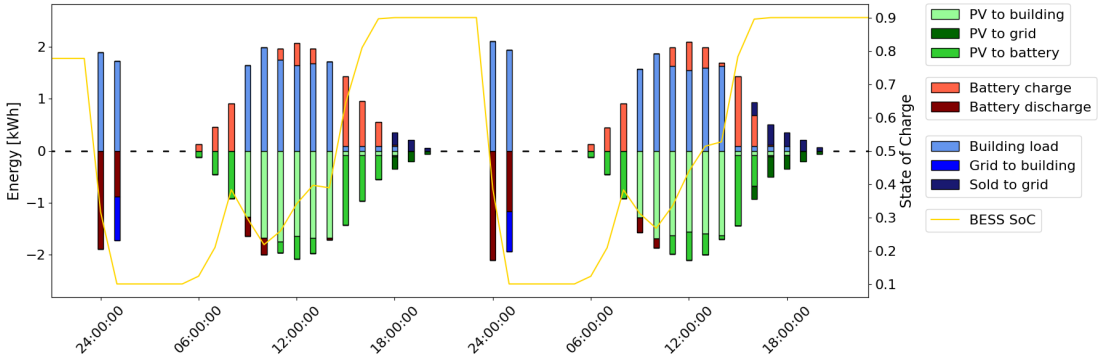

Figure 5.12: Energy flows for all equipment and BESS SoC during two consecutive week-days from 31/06 to 01/07 by adopting RLC at 4800 Wh battery capacity and 10 m^3 storage tank

average daily demand. It is also noted that the PV surplus charges the battery in the morning while in the afternoon it is injected to the grid.

The RL algorithm is expected to try to better match PV production and chiller operation, but also, it seeks to avoid buying expensive electricity and grid injection as much as possible.

Fig. 5.12 and Fig. 5.13 reports the RL control strategy for the same two consecutive days, where the difference with the baseline is highlighted. The energy sold to grid

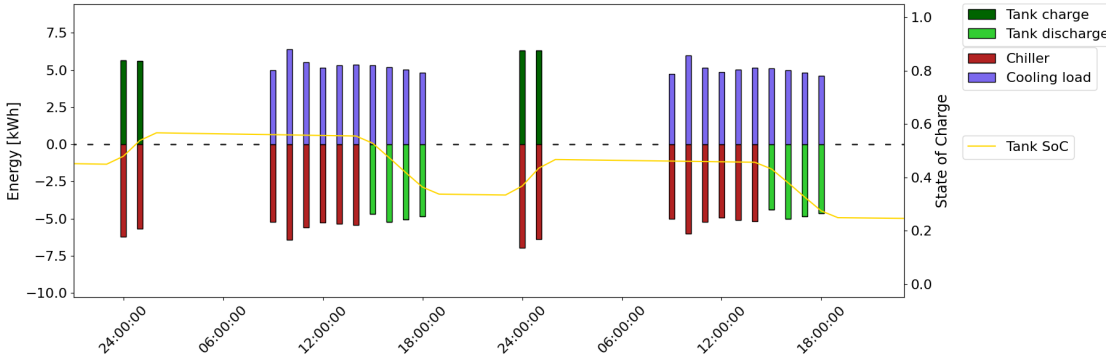

Figure 5.13: Thermal flows for chiller and TES during two consecutive week-days from 31/06 to 01/07 by adopting RLC at 4800 Wh battery capacity and 10 m^3 storage tank

is minimized by the RL by switching on the chiller when the battery is charged and the PV production is still available. The battery is also discharged in order to fill gap between the building demand and the energy supplied by the PV, hence it is avoided to draws energy from the grid during high price hours.

Moreover,the charging phase is shortened since the control agent expects the PV surplus to satisfy the cooling load, resulting in a lower amount of energy exchanged with the main grid. The same control strategy can be seen for each configuration. In terms of net average daily demand and average daily peak power, RL involves notable reductions between 59.5% and 74.3% for the former, and between for the latter. In the case of RB, BESS nominal capacity plays an important role at decreasing both these two quantities, while the TES volume is not influencing the average daily demand and helps reduce average daily peak power. RL has the behaviour as regarding to the BESS size, but TES size has not a clear implication. The electrical duration curve shows more in detail the behaviour of the cooling system. Two duration curves are reported, considering the total building load and the net building load. Concerning the total building load, there is a little difference between RL and RB, except for the cooling system operating time which

is reduced by RL for each configuration. Nonetheless the chiller has basically the same operating time. Differently, the net building load undergoes a reduction of the time the grid is feeding the building, but the net peak power is not affected much, thus the average load power is reduced as well. Moreover, the decrease is much larger and strictly related to the BESS capacity. On average, the grid supply time is 29.4%, 46.1% and 58.1% lower, for 2400 Wh , 4800 Wh and 7200 Wh when using RLC. These results are resumed by Fig. 5.14, Fig. 5.15 and Fig. 5.16.

5.2 Discussions

The previous simulation outputs are discussed in order to fully understand the main implications. The discussion must take into account the stochasticity of the RL algorithm, whose strategy is affected by the intrinsic randomness of the trial and error procedure. Particularly, in some cases, RL unexpectedly decreases its performance when the BESS size is increased, even though the cost rise has low relevance. Nonetheless, the variance of the results does not undercut the confidence, hence the conclusions. The main aim of this work is to assess the importance of implementing an advanced control strategy in the optimization procedure at the design level. Also, it is possible to understand how different sizes of storage technologies affects the performance whichever the control strategy, as well as the improvement by switching to an advanced control strategy, allowing scaling down the storage technologies without affecting the energy cost.

This study focuses on the operational pattern of the case study, so that the best practice for managing storage can be identified and more appropriate and adaptable RBC may be found.

In terms of operating cost, RLC performs better than RBC, since the seasonal energy cost is reduced for all of the combinations, this is also true if the total energy consumption and net energy demand requested to the grid are considered.

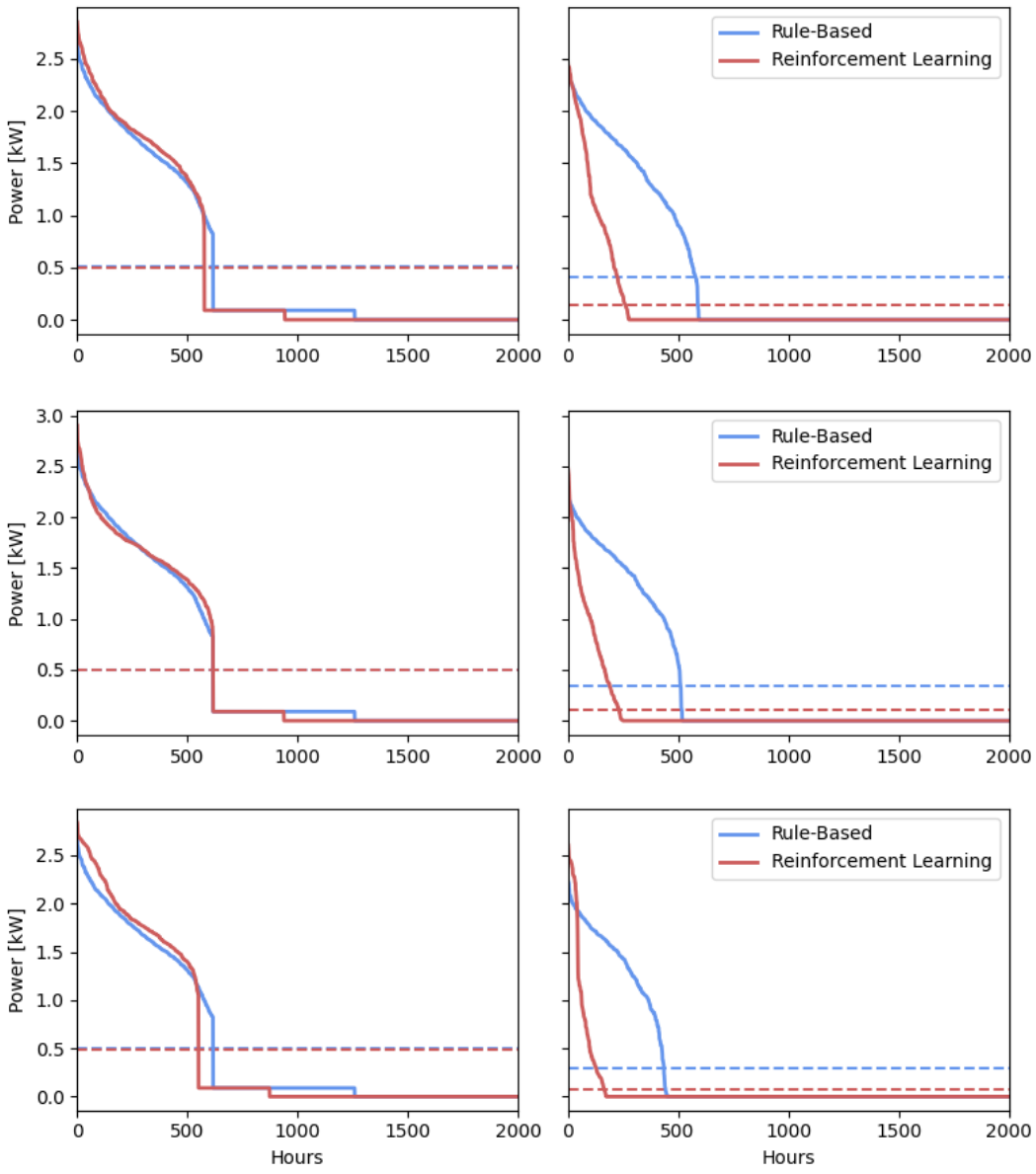

Figure 5.14: Configuration: 1, 2, 3. Duration curve of total (on the left-hand side)and net (on the right-hand side) building load, with their average (dashed line).

RB proved to be very sensitive to the storage size, resulting in a huge impact from the initial design on the electricity cost and particularly from the BESS size. On the other hand, the RL is able to achieve very high economic savings already at

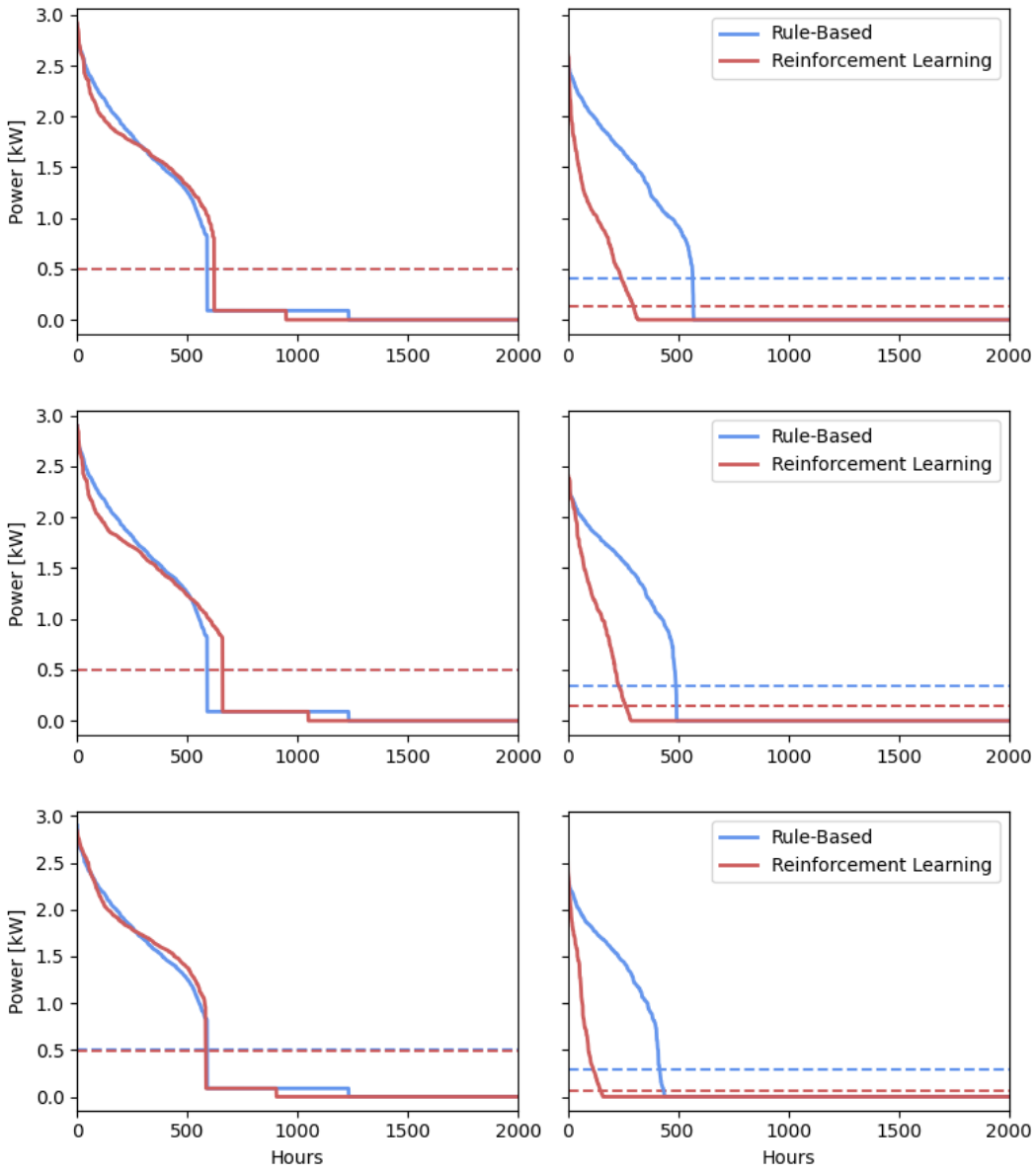

Figure 5.15: Configuration: 4, 5, 6. Duration curve of total (on the left-hand side)and net (on the right-hand side) building load, with their average (dashed line).

low size, but as the storage size increases, the improvement achieved by RL are less intense than those achieved by RB. The better adaptability of RLC appears as its strong point, and this feature is requested more and more by advanced control

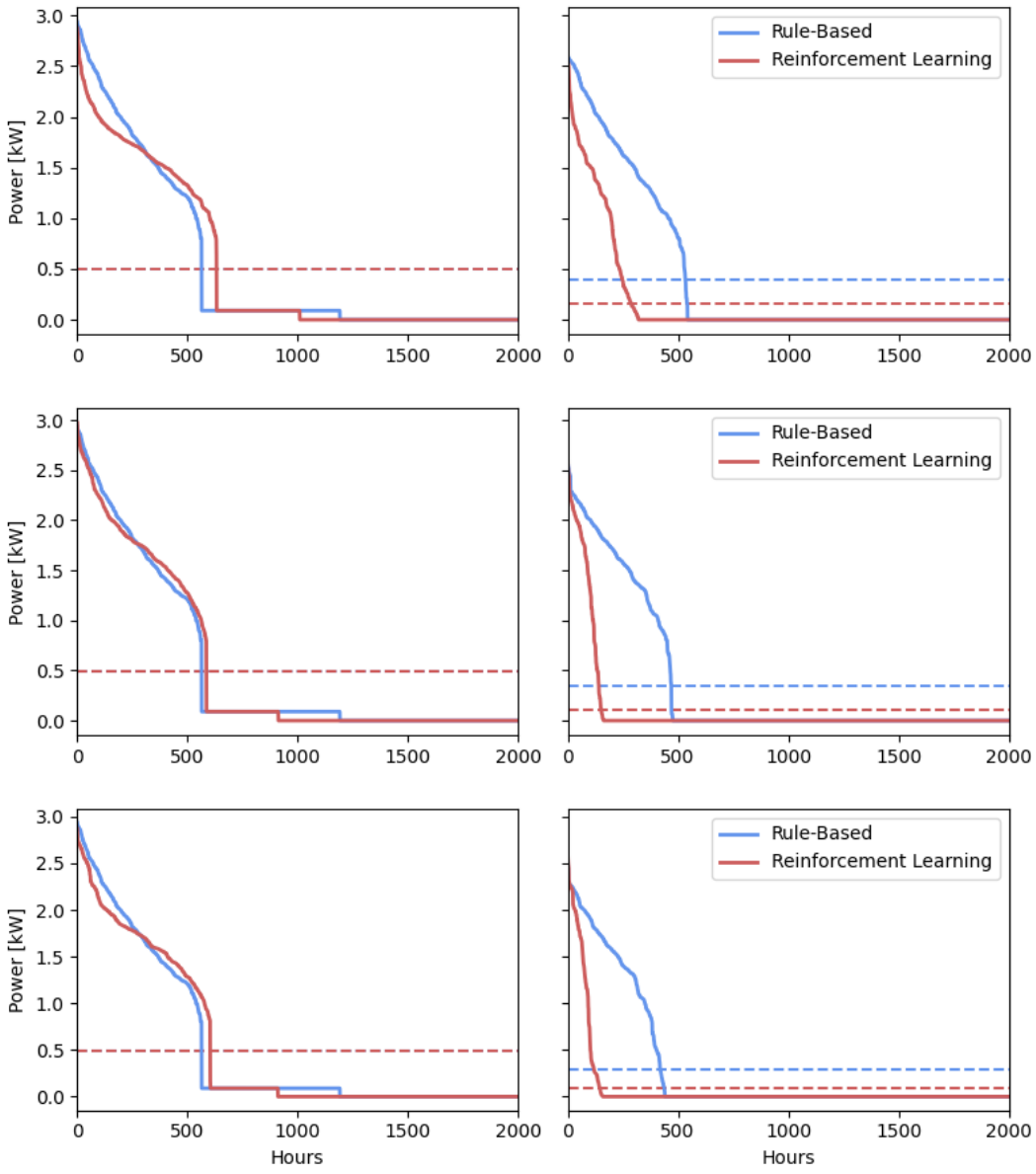

Figure 5.16: Configuration: 7, 8, 9. Duration curve of total (on the left-hand side)and net (on the right-hand side) building load, with their average (dashed line).

strategies.

The configuration that achieves the least operating cost is different for RL and RB.

RL has not shown a clear implication of the storage size on the energy cost, rather

it performs very well at each configuration. Nonetheless, it can be seen that larger tank involves higher thermal losses for both the strategies. Also, the larger BESS helps RL reducing the tank operation, which is not affected when adopting RB.
BESS is largely considered as the best way to increase self-consumption, but the operating cost keeps decreasing as long as the PV injects electricity to the grid, which means there is not room of improvement left for the self-consumption level, beyond that, any larger BESS only comes with higher investment cost. By the time of this work, investment cost is still not sustainable. A preliminary economic analysis shows that the cost savings from higher BESS capacity does not justify the extra investment cost, so that the optimum BESS capacity is different once the investment cost are considered, as it has also been stated by many research. As the BESS capital cost keep decreasing thanks to technical innovation, RL provides a solution to increase levels of self-sufficiency and self-consumption without increasing the size of the battery. This is a massive achievement given that BESS capital cost is one of the most relevant fraction of cost of the energy systems.
The last consideration on the BESS size concerns its role during the operational phase. RL and RB remark two different situations. The capacity is not affecting the energy consumption, hence the chiller and tank schedule, and PV share among the electricity supply, when RB is used. The reason might be that RBC has two distinctive set of rules which include information about either the cooling system either the electrical system. It is supposed that this factor is limiting the ability of the RBC to perform well; oppositely, BESS capacity does have an impact on those quantities when RL is used. RL overcome this problem by making its decision through a set of information comprising those about cooling and electrical system together, and so that, it finds its own set of rules.
The absence of a building thermostatic control does not leave room of improvement in this sense, even though the energy consumption is slightly higher in the case

of RBC because the chiller is never feeding directly the thermal zones, rather it charges the tank, hence the thermal losses increase and so does the electrical demand.

It can be easily assessed that the cost savings are not related to a lower average electricity purchasing cost, but they are due to the lower amount of electricity requested by the grid along with the growth of the on-site consumption. The aim of reducing the energy exchange with the local grid is at the very basis of the use of BESS coupled with PV, which means higher profitability of storage technologies and higher flexibility of the whole energy system. When PV is selling to the grid, the system ends up into poor performance in terms of self-consumption. For this reason, RLC aims at matching PV production and chiller operation as much as possible so that, it also occurs that energy is drawn from the grid during afternoon, thus the specific energy cost is increased. In this way, RL not only avoids that BESS unnecessarily operates, which involves electrical losses due to the round-trip efficiency and converter efficiency, but it reduces the need for high BESS capacity as well, since the excess energy is lower.

The RL algorithm outperforms the baseline at managing the PV-BESS system, even though the control action does not act directly on the charge/discharge of the battery. The level of Self-Sufficiency and Self-Consumption makes the use of electricity storage technologies much more desirable from the point of view of the building flexibility at each size. Moreover, RL does not need large storage utilities to achieve appreciable level of SS and SC when it comes to nZEB design; indeed, it is true that the idea of low-energy building desirably would help easing the burden on the distribution grid, in rural, semi-urban and urban network. This would be a turning point in a system where the number of connections and pro-sumer is drastically increasing. In these terms, advanced and smart control strategies are needed to come into play at the operational level, as suggested by the results

obtained in this work, and at supervisory level. In order to make one thing clear, the adoption of feed-in tariff changes completely the scenario where the events take place, however, this application finds relative importance in a few contexts. High V-RES penetration in the energy system leads to unbalanced grid operation, i.e. duck curve, so that policy makers try to encourage self-consumption or rather to adopt more advanced DR programs like Real-Time Pricing.

Finally, the outlook of the control strategy by RL gives idea on how to develop a more tailored and complex RBC. Machine learning has already been used to find rules for classifying the elements of a given data-set. This time the rules for managing an energy system can be derived. The problem of the RB is that the TES control has only information about the chiller and the tank, also the BESS control only takes into account only the net load as driven parameters. For this reason, it seems advisable to have a single integrated RB strategy fed with information from both the units. RL smart agent makes decision according to the parameters in the observation state. On the basis of the daily pattern provided by the RL agent, a rule for switching on the chiller during the afternoon can be extrapolated. Chiller can feed the thermal zones whenever the PV excess is enough to supply it with electricity and whatever the cost is. In this case, BESS stops charging and eventually, feeds the chiller, which means the TES is switched off and any PV excess can be injected to the battery. An RBC developed from an RLC implementation encloses their advantages providing effectiveness whilst it saves computational effort in the real-time operation, and still it can considered an advanced control strategy, where RL is the tools used to get the smart RBC.

Chapter 6

Conclusions

This study proposes to analyze from the operational point-of-view a multi-energy buildings with on-site electricity generation and storage facilities. The building is located in Turin, and it used by students as it is made up by two study rooms with workstations, a control room and a technical room. It is simulated the period from 1^{st} June to 31^{st} August, which stands for the cooling season. An HVAC system serves the two study rooms and the control room, by means of fun coils. The cooling load is satisfied by an electric chiller that can feed the HVAC system directly or use a cold water storage tank as buffer. The building is connected to the distribution grid where AC is circulating. A PV-BESS system is in charge of producing, storing on-site and feeding electricity to the AC bus through a mono-directional DC/AC inverter.

The weather data comprises the records for outdoor temperature and solar radiation during the aforementioned time, which are provided by the meteorological station of the Caselle airport.

Each room is considered fully occupied during the opening time, that goes from 8:30 a.m. to 6:00 p.m. on the week-days, as it is closed on the week-end, except for the technical room that does not foresee any occupation.

The electricity price follows a Time Of Use tariff scheme with 3 price classes outlining low, medium and high price periods. The selling price is considered constant in time and with the amount of energy sold, and it is set equal to one third of the low price.

The building envelope, along with the cooling system is implemented into EnergyPlus for the simulation, by importing the geometry as a 3D model drawn on SketchUp. The BESS is modeled according to the State Of Charge model whereas the PV production is extrapolated thanks to an empirical model for the selected type available in literature.

The control strategy is in charge of mapping out the operational pattern across all of the occurring states. Two different control strategies are implemented to manage the cold water storage tank, one is a classical Rule-Based Control and the other takes advantage of the Reinforcement Learning methodology. The RBC is fed with the electricity price and the SoC of the tank and returns the charge/discharge state, so that the tank is charged during low price hours in order to maintain a minimum level of charge and it is discharge during peak price hours. The RL control is fed with a set of information of the system concerning both the thermal and electrical side and utilises a smart agent to make decisions. The BESS is managed according to the most widespread RBC available in the literature, which prescribes charging the battery when the PV production exceeds the building electrical load, otherwise discharges. The control problem is so defined by a discrete action space with two options, referring to charge and discharge of the tank, and by the objective function which is the seasonal electricity cost. The problem is also constrained by several technical limitations and by the need of meeting thermal and electrical loads.

These control strategies are implemented for several combinations of BESS and TES capacities to analyze how they can adapt to different situations. The energy model and the control strategy are implemented in a simulation environment based

on the OpenAI Gym from Python library. It is used to initialize and reset the EnergyPlus instance during the whole process. Python is in charge of handling PV and BESS equations as well, while the Building Control Virtual Test Bed (BCVTB), along with the ExternalInterface-Ptolemy server command are used to allow the information exchange between the two softwares. Two more Python classes have been coded to define the RBC agent and the RLC agent for the tank , while sharing the same RBC for the battery.

The RL agent makes use of a Soft Actor-Critic algorithm with discrete action space, implemented thanks to the PyTorch library.

The simulation output are then plotted and presented to be discussed. RLC is affected by a certain degree of stochasticity so that it is convenient to run the algorithm across several seeds and average the results. Nonetheless, RLC proved to be much more efficient than RBC given the same storage capacity and to reduce the operating cost between 5.4% and 25.7%, total energy consumption between 2.0% and 2.6% and net energy demand requested to the grid between 26.0% and 25.2%. RB results in being more sensitive to the storage size, giving a lot of importance to the initial design, whereas RL achieves high economic savings already at low size, and the advantatge with respect to RB narrows down as the size increases. This shows the adaptability of RLC to different situations, which suggests a better flexibility.

Given the absence of thermostatic building, it has not been possible to study the ability of reducing the energy consumption by RLC; indeed the reduction is due to the chiller bypassing the tank.

Once the TES is large enough to supply the cooling load during high price period, there is no point in expand it, which means the average electricity purchasing price can not be reduced anymore, rather the investment cost increases. BESS capacity helps storing the excess PV production, in order to increase the self-consumption.

At the present, high investment cost still prevents from pushing BESS capacity up to the point where all of the excess energy is stored. In this sense, RL notably increases levels of self-sufficiency and self-consumption given the same BESS capacity, and reduces the energy exchanged with the grid as well, which means higher building flexibility, relief to the grid in case of highly penetrated V-RES in the energy system, and BESS capital cost cutting.

Finally, RL provides a recurrent pattern for the operation of the storage utilities. Guidelines for defining more complex RBC can be identified by analyzing the control strategy. First of all, RL shows the relevance of matching PV production and chiller operation. Also, it can act on the BESS cycle by scheduling the chiller operation by receiving information about both thermal and electrical system. These are suggestions to build advanced RBC, where more information is implemented in the decision process and particularly, the building load can be controlled according to the on-site production.

In future works, cooling load prediction from ANN could substitutes the perfect one, which is more consistent with real-time operation. Also, the RL framework was initially meant to be model-free, but the length of the training forces to build models for pre-training before applying on a real building. For this reason, an investigation on the possibility of sharing the model among a cluster of buildings should be carried out, namely transfer learning, or implementing expert knowledge into the replay buffer. These efforts are strongly needed in order to reduce the training phase as much as possible, which is the major barrier for the penetration of RL in the building control state-of-art.

Bibliography

[1] URL: https://ec.europa.eu/clima/policies/eu-climate-action_en (cit. on p. 2).

[2] Weifeng Liu, Warwick J. McKibbin, Adele C. Morris, and Peter J. Wilcoxen. «Global economic and environmental outcomes of the Paris Agreement». In: *Energy Economics* 90 (2020), p. 104838. ISSN: 0140-9883. DOI: https://doi.org/10.1016/j.eneco.2020.104838. URL: https://www.sciencedirect.com/science/article/pii/S014098832030178X (cit. on p. 2).

[3] International Energy Agency. «Key World Energy Statistics». In: (2006). URL: https://www.iea.org/reports/key-world-energy-statistics-2020 (cit. on pp. 2, 4).

[4] U.S. Department of Energy. «International Energy Outlook 2006». In: *Energy Information Administration* (2006). URL: https://www.eia.gov/outlooks/ieo/ (cit. on pp. 2, 4).

[5] Luis Pérez-Lombard, José Ortiz, and Christine Pout. «A review on buildings energy consumption information». In: *Energy and Buildings* 40.3 (2008), pp. 394–398. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2007.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0378778807001016 (cit. on pp. 3, 4).

[6] IEA. *Policy Pathway - Modernising Building Energy Codes 2013*. 2013. URL: https://www.iea.org/reports/policy-pathway-modernising-building-energy-codes-2013 (cit. on pp. 3, 5).

[7] IEA. *Transition to Sustainable Buildings*. 2013. URL: https://www.iea.org/reports/transition-to-sustainable-buildings (cit. on p. 3).

[8] Antoine Levesque, Robert C. Pietzcker, Lavinia Baumstark, Simon De Stercke, Arnulf Grübler, and Gunnar Luderer. «How much energy will buildings consume in 2100? A global perspective within a scenario framework». In: *Energy* 148 (2018), pp. 514–527. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2018.01.139. URL: https://www.sciencedirect.com/science/article/pii/S0360544218301671 (cit. on p. 3).

[9] IPCC. *Intergovernmental Panel on Climate Change. Climate Change 2014: Mitigation of Climate Change*. 2014. Chap. 9: Buildings (cit. on pp. 4, 5).

[10] Xiaoyang Zhong, Mingming Hu, Sebastiaan Deetman, João F.D. Rodrigues, Hai-Xiang Lin, Arnold Tukker, and Paul Behrens. «The evolution and future perspectives of energy intensity in the global building sector 1971–2060». In: *Journal of Cleaner Production* 305 (2021), p. 127098. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2021.127098. URL: https://www.sciencedirect.com/science/article/pii/S0959652621013172 (cit. on p. 4).

[11] Chen Ren and Shi-Jie Cao. «Development and application of linear ventilation and temperature models for indoor environmental prediction and HVAC systems control». In: *Sustainable Cities and Society* 51 (2019), p. 101673. ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2019.101673. URL: https://www.sciencedirect.com/science/article/pii/S221067071931323X (cit. on p. 4).

[12] Rui Jing, Meng Wang, Ruoxi Zhang, Ning Li, and Yingru Zhao. «A study on energy performance of 30 commercial office buildings in Hong Kong». In: *Energy and Buildings* 144 (2017), pp. 117–128. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.03.042. URL: https://www.sciencedirect.com/science/article/pii/S0378778817301123 (cit. on p. 4).

[13] Junqi Wang, Jin Hou, Jianping Chen, Qiming Fu, and Gongsheng Huang. «Data mining approach for improving the optimal control of HVAC systems: An event-driven strategy». In: *Journal of Building Engineering* 39 (2021), p. 102246. ISSN: 2352-7102. DOI: https://doi.org/10.1016/j.jobe.2021.102246. URL: https://www.sciencedirect.com/science/article/pii/S2352710221001029 (cit. on p. 4).

[14] Kui Shan, Shengwei Wang, Dian-ce Gao, and Fu Xiao. «Development and validation of an effective and robust chiller sequence control strategy using data-driven models». In: *Automation in Construction* 65 (2016), pp. 78–85. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2016.01.005. URL: https://www.sciencedirect.com/science/article/pii/S0926580516300097 (cit. on p. 4).

[15] Samuel Thomas and Jan Rosenow. «Drivers of increasing energy consumption in Europe and policy implications». In: *Energy Policy* 137 (2020), p. 111108. ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2019.111108. URL: https://www.sciencedirect.com/science/article/pii/S0301421519306950 (cit. on p. 5).

[16] Heyd F. Más and Dirk Kuiken. «Beyond energy savings: The necessity of optimising smart electricity systems with resource efficiency and coherent waste policy in Europe». In: *Energy Research Social Science* 70 (2020), p. 101658. ISSN: 2214-6296. DOI: https://doi.org/10.1016/j.erss.2020.

101658. URL: https://www.sciencedirect.com/science/article/pii/S2214629620302334 (cit. on pp. 5, 7).

[17] Shady Attia et al. «Overview and future challenges of nearly zero energy buildings (nZEB) design in Southern Europe». In: *Energy and Buildings* 155 (2017), pp. 439–458. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.09.043. URL: https://www.sciencedirect.com/science/article/pii/S0378778817331195 (cit. on p. 6).

[18] «Sustainability Assessment in the Construction Sector: Rating Systems and Rated Buildings». In: *Sustainable Development* 20 (Oct. 2012). DOI: 10.1002/sd.532 (cit. on p. 6).

[19] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, and F. Santos García. «A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect diagnosis». In: *Journal of Building Engineering* 33 (2021), p. 101692. ISSN: 2352-7102. DOI: https://doi.org/10.1016/j.jobe.2020.101692. URL: https://www.sciencedirect.com/science/article/pii/S2352710220310627 (cit. on pp. 6, 7).

[20] B.B. Gupta and Megha Quamara. «An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols». In: *Concurrency and Computation: Practice and Experience* 32.21 (2020). e4946 CPE-18-0159.R1, e4946. DOI: https://doi.org/10.1002/cpe.4946. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4946. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4946 (cit. on p. 6).

[21] Arun Kumar, Sharad Sharma, Nitin Goyal, Aman Singh, Xiaochun Cheng, and Parminder Singh. «Secure and energy-efficient smart building architecture with emerging technology IoT». In: *Computer Communications* 176 (2021), pp. 207–217. ISSN: 0140-3664. DOI: https://doi.org/10.1016/j.comcom.

2021.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0140366421002279 (cit. on pp. 6, 7).

[22] Arun Kumar and Sharad Sharma. «Enhanced Energy-Efficient Heterogeneous Routing Protocols in WSNs for IoT Application». In: Volume 9 (Oct. 2019), pp. 2249–8958. DOI: 10.35940/ijeat.A1342.109119 (cit. on p. 6).

[23] Shanti Pless and Paul Torcellini. «Controlling Capital Costs in High Performance Office Buildings: A Review of Best Practices for Overcoming Cost Barriers». In: (May 2012). URL: https://www.osti.gov/biblio/1043771 (cit. on p. 7).

[24] Paul Torcellini, Shanti Pless, and Matt Leach. «A pathway for net-zero energy buildings: creating a case for zero cost increase». In: *Building Research & Information* 43.1 (2015), pp. 25–33. DOI: 10.1080/09613218.2014.960783. eprint: https://doi.org/10.1080/09613218.2014.960783. URL: https://doi.org/10.1080/09613218.2014.960783 (cit. on p. 7).

[25] Christian Finck, Paul Beagon, John Clauß, Thibault Péan, Pierre Vogler-Finck, Kun Zhang, and Hussain Kazmi. «Review of applied and tested control possibilities for energy flexibility in buildings - A technical report from IEA EBC Annex 67 Energy Flexible Buildings». In: (May 2018). DOI: 10.13140/RG.2.2.28740.73609 (cit. on pp. 7–9).

[26] Zhe Wang and Tianzhen Hong. «Reinforcement learning for building controls: The opportunities and challenges». In: *Applied Energy* 269 (2020), p. 115036. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2020.115036. URL: https://www.sciencedirect.com/science/article/pii/S0306261920305481 (cit. on pp. 8, 23, 25).

[27] Abdul Afram and Farrokh Janabi-Sharifi. «Theory and applications of HVAC control systems – A review of model predictive control (MPC)». In: *Building*

and Environment 72 (Feb. 2014), pp. 343–355. DOI: 10.1016/j.buildenv.2013.11.016 (cit. on pp. 8, 9).

[28] Manfred Morari and Jay H. Lee. «Model predictive control: past, present and future». In: *Computers Chemical Engineering* 23.4 (1999), pp. 667–682. ISSN: 0098-1354. DOI: https://doi.org/10.1016/S0098-1354(98)00301-9. URL: https://www.sciencedirect.com/science/article/pii/S0098135498003019 (cit. on p. 8).

[29] Samuel Prívara, Jan Široký, Lukáš Ferkl, and Jiří Cigler. «Model predictive control of a building heating system: The first experience». In: *Energy and Buildings* 43.2 (2011), pp. 564–572. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2010.10.022. URL: https://www.sciencedirect.com/science/article/pii/S0378778810003749 (cit. on p. 8).

[30] Henrik Karlsson and Carl-Eric Hagentoft. «Application of model based predictive control for water-based floor heating in low energy residential buildings». In: *Building and Environment* 46.3 (2011), pp. 556–569. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2010.08.014. URL: https://www.sciencedirect.com/science/article/pii/S0360132310002672 (cit. on p. 8).

[31] Ion Hazyuk, Christian Ghiaus, and David Penhouet. «Optimal temperature control of intermittently heated buildings using Model Predictive Control: Part II – Control algorithm». In: *Building and Environment* 51 (2012), pp. 388–394. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2011.11.008. URL: https://www.sciencedirect.com/science/article/pii/S0360132311003921 (cit. on p. 8).

[32] Shui Yuan and Ronald Perez. «Multiple-zone ventilation and temperature control of a single-duct VAV system using model predictive strategy». In: *Energy and Buildings* 38.10 (2006), pp. 1248–1261. ISSN: 0378-7788. DOI:

https://doi.org/10.1016/j.enbuild.2006.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0378778806000764 (cit. on p. 8).

[33] Yudong ma, Francesco Borrelli, Hencey B., Andrew Packard, and Scott Bortoff. «Model Predictive Control of Thermal Energy Storage in Building Cooling Systems». In: *Proceedings of the IEEE Conference on Decision and Control* (Jan. 2010), pp. 392–397. DOI: 10.1109/CDC.2009.5400677 (cit. on p. 8).

[34] Benjamin Paris, Julien Eynard, Stéphane Grieu, Thierry Talbert, and Monique Polit. «Heating control schemes for energy management in buildings». In: *Energy and Buildings* 42.10 (2010), pp. 1908–1917. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2010.05.027. URL: https://www.sciencedirect.com/science/article/pii/S0378778810001891 (cit. on p. 9).

[35] Georgios Kontes et al. «Simulation-Based Evaluation and Optimization of Control Strategies in Buildings». In: *Energies* 11 (Dec. 2018), p. 3376. DOI: 10.3390/en11123376 (cit. on p. 9).

[36] Matthew Lai. «Giraffe: Using Deep Reinforcement Learning to Play Chess». In: (Sept. 2015) (cit. on p. 10).

[37] Volodymyr Mnih et al. «Human-level control through deep reinforcement learning». In: *Nature* 518 (Feb. 2015), pp. 529–33. DOI: 10.1038/nature14236 (cit. on p. 10).

[38] David Silver et al. «Mastering the game of Go with deep neural networks and tree search». In: *Nature* 529 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961 (cit. on p. 10).

[39] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. «State-of-the-art on research and applications of machine learning in the building life

cycle». In: *Energy and Buildings* 212 (2020), p. 109831. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2020.109831. URL: https://www.sciencedirect.com/science/article/pii/S0378778819337879 (cit. on p. 10).

[40] Soteris Kalogirou. «Applications of artificial neural-networks for energy systems». In: *Applied Energy* 67 (Sept. 2000), pp. 17–35. DOI: 10.1016/B978-0-08-043877-1.50005-X (cit. on p. 13).

[41] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. «State-of-the-art on research and applications of machine learning in the building life cycle». In: *Energy and Buildings* 212 (2020), p. 109831. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2020.109831. URL: https://www.sciencedirect.com/science/article/pii/S0378778819337879 (cit. on pp. 13, 16).

[42] Simeng Liu and Gregor Henze. «Evaluation of Reinforcement Learning for Optimal Control of Building Active and Passive Thermal Storage Inventory». In: *Journal of Solar Energy Engineering-transactions of The Asme - J SOL ENERGY ENG* 129 (May 2007). DOI: 10.1115/1.2710491 (cit. on p. 13).

[43] Ross May. «The reinforcement learning method: A feasible and sustainable control strategy for efficient occupant-centred building operation in smart cities». PhD thesis. Oct. 2019. DOI: 10.13140/RG.2.2.29921.86883 (cit. on p. 13).

[44] Yuting An, Tongling Xia, Ruoyu You, Dayi Lai, Junjie Liu, and Chun Chen. «A reinforcement learning approach for control of window behavior to reduce indoor PM2.5 concentrations in naturally ventilated buildings». In: *Building and Environment* 200 (2021), p. 107978. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2021.107978. URL: https://www.sciencedirect.com/science/article/pii/S0360132321003826 (cit. on p. 13).

[45] Shunian Qiu, Zhenhai Li, Zhengwei Li, Jiajie Li, Shengping Long, and Xiaoping Li. «Model-free control method based on reinforcement learning for building cooling water systems: Validation by measured data-based simulation». In: *Energy and Buildings* 218 (2020), p. 110055. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2020.110055. URL: https://www.sciencedirect.com/science/article/pii/S0378778819339945 (cit. on p. 14).

[46] Zhen Yu and Arthur Dexter. «Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning». In: *Control Engineering Practice* 18.5 (2010), pp. 532–539. ISSN: 0967-0661. DOI: https://doi.org/10.1016/j.conengprac.2010.01.018. URL: https://www.sciencedirect.com/science/article/pii/S0967066110000353 (cit. on p. 14).

[47] Sha-Xu Zhou and Xian-Fang Li. «Interfacial debonding of an orthotropic half-plane bonded to a rigid foundation». In: *International Journal of Solids and Structures* 161 (2019), pp. 1–10. ISSN: 0020-7683. DOI: https://doi.org/10.1016/j.ijsolstr.2018.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0020768318304347 (cit. on p. 14).

[48] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G.S. Stavrakakis. «Reinforcement learning for energy conservation and comfort in buildings». In: *Building and Environment* 42.7 (2007), pp. 2686–2698. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2006.07.010. URL: https://www.sciencedirect.com/science/article/pii/S0360132306001880 (cit. on pp. 14, 22, 24).

[49] Tim Leurs, Bert J. Claessens, Frederik Ruelens, Sam Weckx, and Geert Deconinck. «Beyond theory: Experimental results of a self-learning air conditioning

unit». In: (2016), pp. 1–6. DOI: 10.1109/ENERGYCON.2016.7513916 (cit. on p. 14).

[50] Frederik Ruelens, Sandro Iacovella, Bert Claessens, and Ronnie Belmans. «Learning Agent for a Heat-Pump Thermostat With a Set-Back Strategy Using Model-Free Reinforcement Learning». In: *Energies* 8 (June 2015). DOI: 10.3390/en8088300 (cit. on p. 14).

[51] Oscar De Somer, Ana Soares, Koen Vanthournout, Fred Spiessens, Tristan Kuijpers, and Koen Vossen. «Using reinforcement learning for demand response of domestic hot water buffers: A real-life demonstration». In: (2017), pp. 1–7. DOI: 10.1109/ISGTEurope.2017.8260152 (cit. on p. 14).

[52] José R. Vázquez-Canteli, Stepan Ulyanin, Jérôme Kämpf, and Zoltán Nagy. «Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities». In: *Sustainable Cities and Society* 45 (2019), pp. 243–257. ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2018.11.021. URL: https://www.sciencedirect.com/science/article/pii/S2210670718314380 (cit. on p. 14).

[53] Silvio Brandi, Marco Savino Piscitelli, Marco Martellacci, and Alfonso Capozzoli. «Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings». In: *Energy and Buildings* 224 (2020), p. 110225. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2020.110225. URL: https://www.sciencedirect.com/science/article/pii/S0378778820308963 (cit. on p. 14).

[54] Young Ran Yoon and Hyeun Jun Moon. «Performance based thermal comfort control (PTCC) using deep reinforcement learning for space cooling». In: *Energy and Buildings* 203 (2019), p. 109420. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2019.109420. URL: https://www.

sciencedirect.com/science/article/pii/S0378778819310692 (cit. on p. 15).

[55] Xianzhong Ding, Wan Du, and Alberto Cerpa. «OCTOPUS: Deep Reinforcement Learning for Holistic Smart Building Control». In: (Nov. 2019), pp. 326–335. DOI: 10.1145/3360322.3360857 (cit. on p. 15).

[56] Anchal Gupta, Youakim Badr, Ashkan Negahban, and Robin G. Qiu. «Energy-efficient heating control for smart buildings with deep reinforcement learning». In: *Journal of Building Engineering* 34 (2021), p. 101739. ISSN: 2352-7102. DOI: https://doi.org/10.1016/j.jobe.2020.101739. URL: https://www.sciencedirect.com/science/article/pii/S2352710220333726 (cit. on p. 15).

[57] Zhanhong Jiang, Michael J. Risbeck, Vish Ramamurti, Sugumar Murugesan, Jaume Amores, Chenlu Zhang, Young M. Lee, and Kirk H. Drees. «Building HVAC control with reinforcement learning for reduction of energy cost and demand charge». In: *Energy and Buildings* 239 (2021), p. 110833. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2021.110833. URL: https://www.sciencedirect.com/science/article/pii/S0378778821001171 (cit. on p. 16).

[58] Bingqing Chen, Zicheng Cai, and Mario Bergés. «Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy». In: *BuildSys '19: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Nov. 2019), pp. 316–325. DOI: 10.1145/3360322.3360849 (cit. on p. 16).

[59] Donald Azuatalam, Wee-Lih Lee, Frits de Nijs, and Ariel Liebman. «Reinforcement learning for whole-building HVAC control and demand response». In: *Energy and AI* 2 (2020), p. 100020. ISSN: 2666-5468. DOI: https://doi.

org/10.1016/j.egyai.2020.100020. URL: https://www.sciencedirect.com/science/article/pii/S2666546820300203 (cit. on p. 16).

[60] Xiaoshun Zhang, Tao Bao, Tao Yu, Bo Yang, and Chuanjia Han. «Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid». In: *Energy* 133 (2017), pp. 348–365. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2017.05.114. URL: https://www.sciencedirect.com/science/article/pii/S036054421730871X (cit. on p. 16).

[61] Dajun Du and Minrui Fei. «A two-layer networked learning control system using actor–critic neural network». In: *Applied Mathematics and Computation* 205 (Nov. 2008), pp. 26–36. DOI: 10.1016/j.amc.2008.05.062 (cit. on p. 16).

[62] Danilo Fuselli, Francesco De Angelis, Matteo Boaro, Stefano Squartini, Qinglai Wei, Derong Liu, and Francesco Piazza. «Action dependent heuristic dynamic programming for home energy resource scheduling». In: *International Journal of Electrical Power Energy Systems* 48 (2013), pp. 148–160. ISSN: 0142-0615. DOI: https://doi.org/10.1016/j.ijepes.2012.11.023. URL: https://www.sciencedirect.com/science/article/pii/S014206151200676X (cit. on p. 17).

[63] Qinglai Wei, Derong Liu, and Guang Shi. «A novel dual iterative Q-learning method for optimal battery management in smart residential environments». In: *IEEE Transactions on Industrial Electronics* 62.4 (2015), pp. 2509–2518. DOI: 10.1109/TIE.2014.2361485 (cit. on p. 17).

[64] Khalid Al-jabery, Zhezhao Xu, Wenjian Yu, Donald C. Wunsch, Jinjun Xiong, and Yiyu Shi. «Demand-Side Management of Domestic Electric Water Heaters Using Approximate Dynamic Programming». In: *IEEE Transactions*

on Computer-Aided Design of Integrated Circuits and Systems 36.5 (2017), pp. 775–788. DOI: 10.1109/TCAD.2016.2598563 (cit. on p. 17).

[65] Shahab Bahrami, Vincent W. S. Wong, and Jianwei Huang. «An Online Learning Algorithm for Demand Response in Smart Grid». In: *IEEE Transactions on Smart Grid* 9.5 (2018), pp. 4712–4725. DOI: 10.1109/TSG.2017.2667599 (cit. on p. 17).

[66] Zhiang Zhang and Khee Lam. «Practical Implementation and Evaluation of Deep Reinforcement Learning Control for a Radiant Heating System». In: (Nov. 2018). DOI: 10.1145/3276774.3276775 (cit. on p. 17).

[67] Zhengbo Zou, Xinran Yu, and Semiha Ergan. «Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network». In: *Building and Environment* 168 (2020), p. 106535. ISSN: 0360-1323. DOI: https://doi.org/10.1016/j.buildenv.2019.106535. URL: https://www.sciencedirect.com/science/article/pii/S0360132319307474 (cit. on p. 17).

[68] June Young Park and Zoltán Nagy. «HVACLearn: A reinforcement learning based occupant-centric control for thermostat set-points». In: June 2020, pp. 434–437. DOI: 10.1145/3396851.3402364 (cit. on p. 17).

[69] Marco Biemann, Fabian Scheller, Xiufeng Liu, and Lizhen Huang. «Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control». In: *Applied Energy* 298 (2021), p. 117164. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2021.117164. URL: https://www.sciencedirect.com/science/article/pii/S0306261921005961 (cit. on p. 17).

[70] Liang Yu, Yi Sun, Zhanbo Xu, Chao Shen, Dong Yue, Tao Jiang, and Xiaohong Guan. *Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings*. June 2020 (cit. on p. 18).

[71] Srinarayana Nagarathinam, Vishnu Menon, Arunchandar Vasan, and Anand Sivasubramaniam. «MARCO - Multi-Agent Reinforcement learning based COntrol of building HVAC systems». In: June 2020, pp. 57–67. DOI: 10.1145/3396851.3397694 (cit. on p. 18).

[72] Ruoxi Jia, Ming Jin, Kaiyu Sun, Tianzhen Hong, and Costas Spanos. «Advanced Building Control via Deep Reinforcement Learning». In: *Energy Procedia* 158 (2019). Innovative Solutions for Energy Transitions, pp. 6158–6163. ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2019.01.494. URL: https://www.sciencedirect.com/science/article/pii/S187661021930517X (cit. on p. 18).

[73] José R. Vázquez-Canteli, Jérôme Kämpf, Gregor Henze, and Zoltan Nagy. «CityLearn v1.0: An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning». In: BuildSys '19 (2019), pp. 356–357. DOI: 10.1145/3360322.3360998. URL: https://doi.org/10.1145/3360322.3360998 (cit. on p. 18).

[74] Giuseppe Pinto, Marco Savino Piscitelli, José Ramón Vázquez-Canteli, Zoltán Nagy, and Alfonso Capozzoli. «Coordinated energy management for a cluster of buildings through deep reinforcement learning». In: *Energy* 229 (2021), p. 120725. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2021.120725. URL: https://www.sciencedirect.com/science/article/pii/S0360544221009737 (cit. on p. 19).

[75] Zhipeng Deng and Qingyan Chen. «Reinforcement learning of occupant behavior model for cross-building transfer learning to various HVAC control systems». In: *Energy and Buildings* 238 (2021), p. 110860. ISSN: 0378-7788.

DOI: https://doi.org/10.1016/j.enbuild.2021.110860. URL: https://www.sciencedirect.com/science/article/pii/S0378778821001444 (cit. on p. 19).

[76] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012. ISBN: 1600490069 (cit. on p. 22).

[77] Y. Wang, Kirubakaran Velswamy, and Biao Huang. «A Long-Short Term Memory Recurrent Neural Network Based Reinforcement Learning Controller for Office Heating Ventilation and Air Conditioning Systems». In: *Processes* 5 (Sept. 2017). DOI: 10.3390/pr5030046 (cit. on p. 23).

[78] Ki Ahn and Cheol-Soo Park. «Application of deep Q-networks for model-free optimal control balancing between different HVAC systems». In: *Science and Technology for the Built Environment* 26 (Oct. 2019), pp. 1–16. DOI: 10.1080/23744731.2019.1680234 (cit. on p. 25).

[79] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018. ISBN: 0262039249 (cit. on p. 26).

[80] Arun Nair et al. «Massively Parallel Methods for Deep Reinforcement Learning». In: (July 2015) (cit. on p. 26).

[81] Hado Van Hasselt, Arthur Guez, and David Silver. «Deep Reinforcement Learning with Double Q-learning». In: (Sept. 2015) (cit. on p. 27).

[82] Tuomas Haarnoja et al. «Soft Actor-Critic Algorithms and Applications». In: (Dec. 2018) (cit. on pp. 28–30).

[83] Giuseppe Pinto, Silvio Brandi, Alfonso Capozzoli, José Vázquez-Canteli, and Zoltán Nagy. «Towards Coordinated Energy Management in Buildings using Deep Reinforcement Learning». In: (Sept. 2020) (cit. on p. 30).

[84] Petros Christodoulou. *Soft Actor-Critic for Discrete Action Settings*. 2019. arXiv: 1910.07207 [cs.LG] (cit. on p. 30).

[85] Angela Amato, Matteo Bilardo, Enrico Fabrizio, Valentina Serra, and F. Spertino. «Energy Evaluation of a PV-Based Test Facility for Assessing Future Self-Sufficient Buildings». In: *Energies* 14 (Jan. 2021), p. 329. DOI: 10.3390/en14020329 (cit. on pp. 33, 38, 41).

[86] Tiansong Cui, Shuang Chen, Yanzhi Wang, Qi Zhu, Shahin Nazarian, and Massoud Pedram. «An optimal energy co-scheduling framework for smart buildings». In: *Integration* 58 (2017), pp. 528–537. ISSN: 0167-9260. DOI: https://doi.org/10.1016/j.vlsi.2016.10.009. URL: https://www.sciencedirect.com/science/article/pii/S0167926016300864 (cit. on p. 37).

[87] Reino Ruusu, Sunliang Cao, Benjamin Manrique Delgado, and Ala Hasan. «Direct quantification of multiple-source energy flexibility in a residential building using a new model predictive high-level controller». In: *Energy Conversion and Management* 180 (2019), pp. 1109–1128. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2018.11.026. URL: https://www.sciencedirect.com/science/article/pii/S0196890418312706 (cit. on p. 38).

[88] Wilhelm Durisch, Bernd Bitnar, Jean-C. Mayor, Helmut Kiess, King-hang Lam, and Josie Close. «Efficiency model for photovoltaic modules and demonstration of its application to energy yield estimation». In: *Solar Energy Materials and Solar Cells* 91.1 (2007), pp. 79–84. ISSN: 0927-0248. DOI: https://doi.org/10.1016/j.solmat.2006.05.011. URL: https://www.sciencedirect.com/science/article/pii/S0927024806003345 (cit. on p. 44).

[89] Mark Z. Jacobson and Vijaysinh Jadhav. «World estimates of PV optimal tilt angles and ratios of sunlight incident upon tilted and tracked PV panels relative to horizontal panels». In: *Solar Energy* 169 (2018), pp. 55–66. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2018.04.030. URL: https://www.sciencedirect.com/science/article/pii/S0038092X1830375X (cit. on p. 44).

[90] URL: https://www.gse.it/servizi-per-te/fotovoltaico/scambio-sul-posto (cit. on p. 46).

[91] Sergio B. Sepúlveda-Mora and Steven Hegedus. «Making the case for time-of-use electric rates to boost the value of battery storage in commercial buildings with grid connected PV systems». In: *Energy* 218 (2021), p. 119447. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2020.119447. URL: https://www.sciencedirect.com/science/article/pii/S0360544220325548 (cit. on p. 53).

[92] Rafael Hirschburger and Anke Weidlich. «Profitability of photovoltaic and battery systems on municipal buildings». In: *Renewable Energy* 153 (2020), pp. 1163–1173. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2020.02.077. URL: https://www.sciencedirect.com/science/article/pii/S096014812030272X (cit. on p. 53).

[93] Pietro Elia Campana, Luca Cioccolanti, Baptiste François, Jakub Jurasz, Yang Zhang, Maria Varini, Bengt Stridh, and Jinyue Yan. «Li-ion batteries for peak shaving, price arbitrage, and photovoltaic self-consumption in commercial buildings: A Monte Carlo Analysis». In: *Energy Conversion and Management* 234 (2021), p. 113889. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2021.113889. URL: https://www.sciencedirect.com/science/article/pii/S0196890421000662 (cit. on p. 53).

[94] Maria M. Symeonidou, Chrysanthi Zioga, and Agis M. Papadopoulos. «Life cycle cost optimization analysis of battery storage system for residential photovoltaic panels». In: *Journal of Cleaner Production* 309 (2021), p. 127234. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2021.127234. URL: https://www.sciencedirect.com/science/article/pii/S0959652621014530 (cit. on p. 53).