A Reinforced Learning (RL) algorithm for optimal control of HVAC: Comparison with a standard PI control for the supermarket DOE reference building

Reinforcement Learning Control Algorithm for HVAC Retrofitting: Application to a Supermarket Building Model by Dynamic Simulation

Antonio Mastropietro[1,3], Fabio Castiglione[2], Stefano Ballesio[1], Enrico Fabrizio[2]
[1] Addfor S.p.A., Torino, Italy
[2] Energy Department, TEBE Research Group, Politecnico di Torino, Torino, Italy
[3] Department of Mathematical Sciences, Politecnico di Torino, Torino, Italy

Abstract

Efficient control of Heating, Ventilation and Air Conditioning systems can lead to great reduction in energy consumption. This can be achieved by new data-driven control algorithms based on Reinforcement Learning (RL). In this work Dynamic Simulation is coupled with a model-free RL algorithm to study its performance in terms of energy saving and thermal comfort in a realistic scenario. Two models are derived from the DOE Supermarket Reference Building for two climate locations. The simulations performed show a reduction between 5.4% and 9.4% in primary energy consumption for the two locations, guaranteeing the same thermal comfort of state-of-the-art controls.

Introduction

The ecological concern about global warming and the need of energy savings has become increasingly present in recent years. The amount of global energy consumption due to buildings is estimated in 40%, according to Prez-Lombard et al. (2008). The reduction of energy consumption and operating cost of existing buildings can be achieved by retrofitting. Instead of acting on the building envelope, retrofit could be implemented by improving the efficiency of the Heating, Ventilation, Air Conditioning (HVAC) systems. The control system of the thermal plant represents a great opportunity to increase the HVAC efficiency. This work explores by simulation a retrofitting scenario of the control system of the HVAC plant on a commercial building, with a view to design a scalable solution on real sites.

The traditional controls for HVAC comprehend the classes of model-based and rule-based controls. Model-based controls like Model Predictive Control can fit to complex thermodynamics and they achieve good results in terms of primary energy saving on a single building. Nevertheless, the retrofit application of model-based methods requires to develop and validate a thermo-energetic model for each existing building. The performance of model-based controls is strictly related to the model quality. An accurate model can be expensive and complex to be obtained at a sufficient level of fidelity in common working conditions. Therefore, the model-based approach is characterized by a high initial investment, an obstacle to the deployment on everyday building projects and to scale up this solution on a great number of buildings (Sturzenegger et al., 2016). Furthermore, when the building envelope undergoes interventions of energy efficiency or the building itself change in the designated use, the model has to be rebuilt and tuned, with an expensive involvement of a domain expert. Conversely, rule-based controls (RBC) are state-of-the-art model-free controls that represents an industry standard. A model-free solution can potentially scale up, because the absence of a model makes the solution easily applicable on different buildings without the need for a domain expert (Ruelens et al., 2015). The main drawback of RBC is that they are difficult to be optimally tuned because they are not enough adaptable with respect to the intrinsic complexity of the coupled building and plant thermodynamics.

Recently a new class of controls have been proposed, composed by data-driven machine learning (ML) algorithms. Among the others, the data-driven algorithms based on Reinforcement Learning (RL) showed promising results as HVAC controls in the recent works of Ruelens et al. (2015) and Wei et al. (2017). RL algorithms (Sutton and Barto, 1998) are suitable to solve optimal control problems in stochastic environments. A lot of applications of RL to real world scenarios has been studied, for example to the battery management systems of electric vehicles, photovoltaic energy storage and domestic hot water (Vázquez-Canteli and Nagy, 2019). The advantage of the RL class over traditional controls is that it provides both model-free and adaptive controls: an RL algorithm interacts directly with the HVAC control system and adapt continuously to the controlled environment using real-time data collected on site without the need to access to a thermo-energetic model of the building. Hence, an RL solution could obtain primary energy saving with respect to a RBC, reducing the operating cost while remaining suitable to a large-scale application.

https://doi.org/10.26868/25222708.2019.210614

Since an RL control learns online by interacting with an environment, simulation tools are very attracting to study an algorithm in the field of building energy management systems. In this work a simulation-based comparison between an RL algorithm and a RBC is presented, using co-simulation with EnergyPlus software. The numerical study is useful to compare the two controls in a reproducible and fair environment. The aim of the analysis presented is to assess the adaptability of a RL control to different operating conditions, thus checking if the same algorithm can scale up to different buildings.

To reduce the cost of the implementation of the RL solution on a real scenario, it is assumed that the data available to the RL control are those commonly used to monitor an HVAC system. The implementation of an RL method requires a small initial investment, comparable to the connection of a single core computer with the HVAC system. Therefore it is attractive for a large-scale application on HVAC systems where the operating cost is high, like those in charge of the thermo-regulation of a great volume. One of the building use class where it could be convenient to implement an RL solution is the supermarket class. Supermarkets are widespread buildings with variable thermal loads and complex occupational patterns that introduce a non-negligible stochastic component from the HVAC control point of view. The variability of those environments is manifest in the retail area, which is characterized by a great volume and it does not require strict comfort requirements. Consequently, it could be interesting to conduct the comparison for the retail area of a supermarket.

The U.S. Department of Energy (DOE) provides validated EnergyPlus Reference Buildings suitable for realistic retrofit applications. Starting from the Supermarket Reference Building, two different building models has been developed within EnergyPlus for Italian locations, using traditional design handbooks (Stefanutti, 2001). During each simulation the control regulates the set-point temperature of the HVAC plant in charge of the thermo-regulation of the Sales zone of the building model, which represents the retail area of the supermarket.

Other attempts of using RL as control strategy for HVAC have already been experimented in a simulated environment but extremely simplified models of buildings and HVAC plants have been used (Barrett and Linder, 2015; Ruelens et al., 2015). To the best of our knowledge, the first attempt to use dynamic simulation coupled with a recent RL algorithm has been described by Wei et al. (2017) but the results obtained are difficult to be reproduced since no building model description is given. The plant considered by Wei et al. (2017) is a variable air flow volume system while in this study a more common case of a constant air flow volume system is studied. Here, inspired by the work of Ruelens et al. (2015), a modified version of the Fitted-Q-Iteration algorithm is applied to the two supermarket models. The first paragraph *Modelling Methods* presents the analysed buildings and it describes the differences with respect to the original DOE model for both climate zone B and E. Then the RBC used as baseline and the RL algorithm for HVAC control are described in detail in the paragraph *Control Methods*. The simulation framework that combines the EnergyPlus Software with the RL training algorithm is presented in paragraph *Simulation Framework*. The simulation results are presented in paragraph *Results* to analyze the performance of the proposed algorithm with respect to the RBC. Finally, conclusions and possibilities of future work are presented in paragraph *Conclusions*.

Table 1: U-Values envelope $[\frac{W}{m^2K}]$.

	Climate Zone E	Climate Zone B
External Wall	0.278	0.429
Ceiling	0.257	0.428
Floor	0.238	0.333
Glazing	1.6	2.6

Modelling methods

The building model has been located in the Italian locations of Bergamo and Catania. Following Koppen climate zone classification, Bergamo is characterized by a climate type *3A* with respectively 778°C and 1099°C cooling and heating Degree-Days (DD); Catania is classified *4A*, with respectively 324°C and 2550°C cooling and heating DD. According to ASHRAE Standards 90.1-2004 and 90.2-2004 climate zone, both the locations are classified as *Cfa*. The Italian territory is divided into six climate zones increasing from A to F as a function of heating DD (D.P.R., 1993). Bergamo belongs to the Italian climate zone *E*, while Catania to the Italian climate zone *B*. In what follows, the locations are referred according to the Italian climate zone classification. Real weather data of two years of 2016-2017 has been collected from open-data sources (ARPA, 2018), to compare the controls in a realistic scenario.

The DOE Supermarket Reference Building is composed by 6 zones, specifically named Bakery, Deli, Dry Storage, Office, Produce and Sales. The biggest area is the Sales zone, which extends for 2325 m^2. In the developed building models each zone is independently thermo-regulated by an RBC which controls the internal air set-point. The RBC is followed by a PI logic that chooses the supply air temperature of the controlled zone. The building model and the PI are tuned to accomplish italian law requirements (MISE, 2015). In this work, the comparison between the controls involve to replace only the RBC of the Sales zone with an RL algorithm while retaining the PI logic to choose the supply air temperature of the controlled zone. The original constructions have been

Table 2: Design set-point (for the Sales zone, the design set-point is substituted by the RBC or RL control), nominal capacities, lightning and electric load equipment, ventilation requirements, design occupancy and activity level (MET = Metabolic Equivalent of Task) for each climate zone and thermal zone.*

	Zone		Bakery	Deli	Dry Storage	Office	Produce	Sales
Set-Point [°C]	B, E	Heating	16	16	14	20	16	(*)
		Cooling	25	25	-	25	25	(*)
Nominal Capacities [kW]	E	Cooling	40	40	32	5.6	28	80
		Heating	32	40	64	12	96	400
	B	Cooling	45	90	65	9	63	225
		Heating	30	30	65	10	90	300
Electrical [$\frac{W}{m^2}$]	B, E	Lightning	15	15	10	15	15	15
		Equipment	15	15	3.5	10	15	3.5
Outdoor Air [$\frac{m^3}{s}$]	B, E		3.4592	1.4607	4.0423	0.0586	0.4620	15.1125
Occupancy [people]	B, E		3	3	62	5	3	465 (max)
MET	B, E		1.8	2.0	2.0	1.0	2.0	1.7

modified to accomplish the different regulations for the two Italian climate zones (MISE, 2008). Table 1 resumes the thermal transmittance (U-value) of constructions for each climate zone.

The number of people of the DOE model has been adapted according to the Italian standard (UNI, 1995). The random occupation model assumed is defined by Page et al. (2008). The parameter of mobility μ is set to 0.5, corresponding to medium mobility. The occupation fraction mean value has been kept as in the original DOE model, between 0.2 in the morning to 0.8 at 4 p.m. in weekdays. Random occupation model has been applied only to the Sales zone while the original occupation schedules have been kept unchanged for all the other zones. The occupancy activity level are set according to the Engineering manual of EnergyPlus, depending on the different activities conducted in each zone (Table 2). The workers of the Bakery and Deli zones are assumed to keep a constant clothing along the year, equal to 0.5 clo, since those zones have high internal gains. The clothing model for the occupants of the other zones is the Dynamic Clothing Model (ASHRAE proposal). The lighting and electric load equipments have been changed to the Italian standard values (UNI, 2017), as resumed in Table 2. Ventilation requirements are specified in Table 2 in terms of outdoor air. Opening hours are from 9 a.m. to 9 p.m. on working days and 9 a.m. to 8 p.m. on each Sunday. The winter season is defined between the 15th October and the 15th April for climate zone E while for climate zone B it is defined from the 1st December to the 31th March (D.P.R., 1993). Heating is available only during winter, while HVAC is in cooling mode during summer.

Control methods

The HVAC system for the Supermarket model is an all-air system: both ventilation and air conditioning are provided by supply air. The HVAC system has six different air loops, one for each thermal zone, that handle the outdoor air flow. The logic of each air loop imposes the delivery air temperature set-point to guarantee the objective air temperature defined by the zone set-point. Each loop is driven by a fan with constant velocity and air diffusers as terminals to the zone. Heating load is supplied by Furnaces and cooling load is supplied by Direct Expansion Chillers, one for each Air Handling Unit. The nominal capacities of the equipments for climate zone B and E obtained by EnergyPlus design simulation are reported in Table 2. The thermostat for the HVAC system provides a single set-point and its logic is described by Stefanutti (2001). Both for climate zone B and E, the values of the design set-point of the five thermal zones of the Supermarket Reference Building, all except the Sales area controlled, are represented in Table 2. These thermal comfort requirements have been chosen from traditional italian manuals depending on the zone destination (Stefanutti, 2001). In the controlled zone, the thermostat set-point is set to keep an objective air temperature above of 16°C in winter and below 25°C in summer.

The state-of-the-art RBC for all-air systems imposes a constant set-point while the Proportional-Integral (PI) control modifies the delivery air temperature set-point (ASHRAE, 2017). The error signal is computed between the RBC thermostat set-point and the actual internal air temperature. The PI control system for both the heating and cooling have been tuned with Ultimate Oscillation method described in ASHRAE (2017). The tuning results of the PI control parameters K_p and K_i are listed in Table 3. This control is established as a baseline for the Sales zone and it will be compared with a RL based control.

Reinforcement Learning

RL is an Artificial Intelligence framework in which an *agent* interacts with an *environment* and learns the optimal sequence of actions, represented by a policy, to reach a desired goal (Sutton and Barto, 1998). In this work, the environment is one of the supermarket building models described before. The learning goal of the agent is expressed by means of a reward,

Table 3: PI tuning for the Climate Zone E and B.

	Climate zone E				Climate zone B			
	Heating		Cooling		Heating		Cooling	
Thermal Zone	K_p	K_i	K_p	K_i	K_p	K_i	K_p	K_i
Bakery	1.802	0.217	2.8	0.3	1.802	0.217	2.8	0.3
Deli	1.802	0.217	2.8	0.3	1.802	0.217	1.5	0.3
Dry Storage	1.802	0.217	1.802	0.217	1.802	0.217	1.802	0.217
Office	1.802	0.217	1.802	0.217	1.802	0.217	1.802	0.217
Produce	1.802	0.217	1.6	0.3	1.5	0.3	1.6	0.3
Sales	1.802	0.217	1.6	0.3	1.802	0.217	2.8	0.3

a scalar feedback value returned to the agent which measures how the agent is behaving with respect to the learning objective. The agent-environment interaction is formalized by means of a Markov Decision Process (MDP), which is completely described by a tuple $(S, A, r, \mathcal{P})$, where S is the set of states, A is the set of actions, $r : S \times A \to \mathbb{R}$ is the reward function and $\mathcal{P}$ is the transition probability from a state-action pair to the next state $\mathcal{P} : S \times A \times S \to [0, 1]$. During a discrete sequence of time-step, the agent improves its policy $\pi : S \to \mathbb{P}(A)$, which represents its behaviour in the environment, where $\mathbb{P}(A)$ is a probability distribution among the global set of actions. The objective of the policy improvement procedure is to find the optimal policy π_*, defined as the policy that maximize the *expected discounted return* $\mathbb{E}[g_t] = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k}]$, where $\gamma \in [0, 1]$.

The core assumption to model the RL framework as a MDP is that the state observed by the agent completely represents the useful information about the environment. In a real application, the Markovian assumption is relaxed in favor of a "quasi-Markovian" condition because the agent can not observe the complete representation of the environment: the state observed by the agent approximates the actual state (Sutton and Barto, 1998). To apply the RL framework to a real environment, the quasi-Markovian approximation is not a negligible assumption: *the signals that compose the observed state have to be carefully chosen*. This work follows the approach of Ruelens et al. (2015), where the observed state is an array composed of past and current data. A similar solution has been proposed by Zhang et al. (2018), where the last 3 historical values of each signal is included in the observed state. Conversely, the arguments about the quasi-Markovian state are discarded in Wei et al. (2017), where it is assumed that the future internal air temperature dynamics is completely determined by actual signals about time, solar radiation, outdoor air temperature and indoor air temperature. The state observed by the RL agent studied in this work is reported in Table 4.

Table 4: State observed by the agent during the RL time-step t, where (t) actual time-step; (t±4) 4 time-step before/forecast; T_{in} *zone internal air temperature of the zone;* T_{send} *delivery air temperature.*

Signal Name	Signals observed
Season, Weekday time-step of the day	t
Electric cost, Solar radiance, Gas cost	$t, t-4$
External air temperature	$t, t+4$
T_{in} Bakery, T_{in} Deli T_{in} Dry Storage, T_{in} Office T_{in} Produce, T_{in} Sales	$t, t-4$
Humidity Sales, T_{send} Sales	$t, t-4$

The objective of the RL algorithm applied to the HVAC control system is to save energy while satisfying comfort constraints. The translation of the learning objective into the reward function has to take into account a trade-off between the two components of comfort and primary energy consumed. In this work the comfort constraint is defined as an acceptable interval on the internal air temperature $[\underline{T_b}, \overline{T_b}]$, where $\underline{T_b}$ and $\overline{T_b}$ depend on the season of the actual time-step. The interval $[\underline{T_b}, \overline{T_b}]$ is fixed to $[16°C, 19°C]$ for winter and $[23°C, 25°C]$ for summer (Stefanutti, 2001). Here, the reward at each time-step r_t is defined as the sum $r_t = r_A + r_B$ of the two components of cost r_A and comfort r_B:

$$\begin{aligned} r_A &:= -\lambda c \\ r_B &:= \begin{cases} 0 & \text{if } T_{in} \in [\underline{T_b}, \overline{T_b}] \\ -\exp(p) & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

where c is the sum of the electric energy and thermal energy costs at each time-step, $p = \max(\underline{T_b} - T_{in}, 0) + \max(T_{in} - \overline{T_b}, 0)$ is the constraint penalization factor and λ is a trade-off parameter between the comfort and the cost component. The exponential function applied on p accounts for the greater importance of the comfort component as the temperature exceed the comfort constraints.

In a real application, it needs to be established a method to evaluate the comfort in the controlled zone to tune the λ trade-off parameter. A tuning criterion can be defined by regarding the comfort constraint as a *stochastic constraint*. The idea is that given a temperature T and an exceeding value ΔT, the agent can exceed the temperature $T + \Delta T$ only with a bounded probability for each hour of the day. Formally, to define the stochastic constraint, first compute the empirical probability density function (empirical pdf) of the internal air temperature values T_{in} registered for each hour of the day. An example of such densities

Table 5: stochastic constraint for λ parameter tuning divided by each season.

Winter	Summer
$mE_{<16} < 20\%$	$mE_{>25} < 20\%$
$mE_{<15.5} < 8\%$	$mE_{>25.5} < 8\%$
$mE_{<15} < 3\%$	$mE_{>26} < 3\%$

for different control systems is showed in Figure 1. For each hour h of the day, let E^h be the probability that T_{in} exceeds the lower temperature $\underline{T_b}$ for winter and the upper temperature $\overline{T_b}$ for summer. The stochastic constraint is defined in terms of the maximum exceeding probability over all the hours of the day, $mE_{<16}$ for winter and $mE_{>25}$ for summer. In addition, $mE_{<15.5}$, $mE_{<15}$ for winter and $mE_{>25.5}$, $mE_{>26}$ for summer can be defined. The set of conditions assumed for the stochastic constraint are reported in Table 5.

The RL algorithm improves continuously its policy by collecting information about the environment dynamics. The information collected are summarized by a set of transitions, that is, a set of tuples $\mathcal{T} = \{(s_t, a_t, r_t, s_{t+1})\}_t$. The expected value of the return obtained by starting in a state s, choosing an action a and following a policy π is called *q-function* q^π for the policy π: $q^\pi(s_t, a_t) = \mathbb{E}_\pi\big[g_t \mid s_t, a_t\big]$. The agent's Q-function Q_π is the approximation of q^π in every pair (s, a). The optimal policy π_* satisfies the Bellman optimality equation:

$$q_*(s_t, a_t) = \mathbb{E}_{\pi_*}\left[r_t + \gamma \max_{a_{t+1}} q_*(s_{t+1}, a_{t+1}) - q_*(s_t, a_t)\right] \tag{2}$$

From equation (2) descends the Q-learning algorithm, first proposed by Watkins (1989): after a transition (s_t, a_t, r_t, s_{t+1}), the Q-function is updated by means of the equation (3):

$$\begin{aligned} \sigma_t^{(1)} &= r_t + \gamma \max_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1}) \\ Q_\pi(s_t, a_t) &\leftarrow (1-\alpha)\, Q_\pi(s_t, a_t) + \alpha\, \sigma_t^{(1)}. \end{aligned} \tag{3}$$

where $\alpha \in [0, 1]$ is the learning-rate. Multi-step methods are introduced to improve the Q-function approximation of Q-learning. (Sutton and Barto, 1998). After having observed K consecutive transitions, the Multi-step Q-learning algorithm is characterized by the following update formula:

$$\begin{aligned} \sigma_t^{(K)} &= \sum_{k=0}^{K-1} \gamma^k r_{t+k} + \gamma^K \max_{a_{t+K}} Q_\pi(s_{t+K}, a_{t+K}) \\ Q_\pi(s_t, a_t) &\leftarrow (1-\alpha)\, Q_\pi(s_t, a_t) + \alpha\, \sigma_t^{(K)}. \end{aligned} \tag{4}$$

Algorithm 1.1: Fitted Q-Iteration algorithm

```
Initialize $\hat{Q}_0 = 0$; day $N = 0$, transitions set $B \leftarrow \emptyset$.
for day $N \leq N_{max}$ do
    for time-step $t$ of day $N$ do
        Acquire transition $b_{t,N}$ using $\hat{Q}_N$;
        $B$: Append $b_{t,N}$
    end
    Set Input $\mathcal{I} \leftarrow \emptyset$; Response $\mathcal{R} \leftarrow \emptyset$;
    for transition $b_j \in B$ do
        Compute $y_j$ relative to $b_j$ with equation (5);
        $\mathcal{I}$: Append $(s_j, a_j)$;
        $\mathcal{R}$: Append $y_j$;
    end
    Train $\hat{Q}_N$ on the input-response pair: $(\mathcal{I}, \mathcal{R})$
end
Outputs a $\hat{Q}_{Nmax}$ approximator.
```

In practice, the application of RL to HVAC control needs a non-linear regression algorithm because of the great dimension of the state space, together with the complexity and non-linearity of the building dynamics. The algorithm analysed in this work is *Fitted Q-Iteration* (FQI), developed firstly by Ernst et al. (2005), where the regression is computed by means of the Extremely Randomized Tree (Geurts et al., 2006). FQI is called a *batch-mode* RL algorithm, because it processes a transitions set in batch. FQI assumes that a buffer of past transitions $B = \{(s_j, a_j, r_j, s_{j+1})\}_{j \leq t}$ has been collected, where t is the current time-step. Then it iterates the regression to the set of transitions in order to improve the accuracy of the Q-function approximation $\hat{Q}$. For each transition $b_j \in B$, the regression procedure uses a response variable value y_j. While in the original FQI the response y_j is derived from Q-learning, in this work y_j is derived from Multi-Step Q-learning:

$$y_j = \sum_{k=0}^{K-1} \gamma^k r_{j+k} + \gamma^K \max_{a_{j+K}} \hat{Q}_{N-1}(s_{j+K}, a_{j+K}) \tag{5}$$

where $\hat{Q}_{N-1}$ is the approximator obtained at the preceding step of the iteration. Finally, the $\hat{Q}_N$ approximator is used to derive a new policy and acquire new transitions. The variant of FQI is called Multi-Step FQI, or MS-FQI, and the training pseudo-code is summarized in box 1.1. The MS-FQI is coupled with is the ϵ-greedy policy in order to explore the environment (Sutton and Barto, 1998). This policy requires the agent to choose with probability $1 - \epsilon$ the action with the maximum estimated Q-value and with probability ϵ a random action.

Simulation framework

The complete simulation framework comprises the building model, the communication interface and the RL framework. The two supermarket models has been developed using EnergyPlus 8.5. The communication interface is implemented by the Building Controls Virtual Test Bed 1.5.0, abbreviated BCVTB (Wetter, 2011). The RL framework is made up of the environment and the agent components.

The environment is compatible with the OpenAI Gym framework (Brockman et al., 2016). After receiving the simulated physical signals from the En-

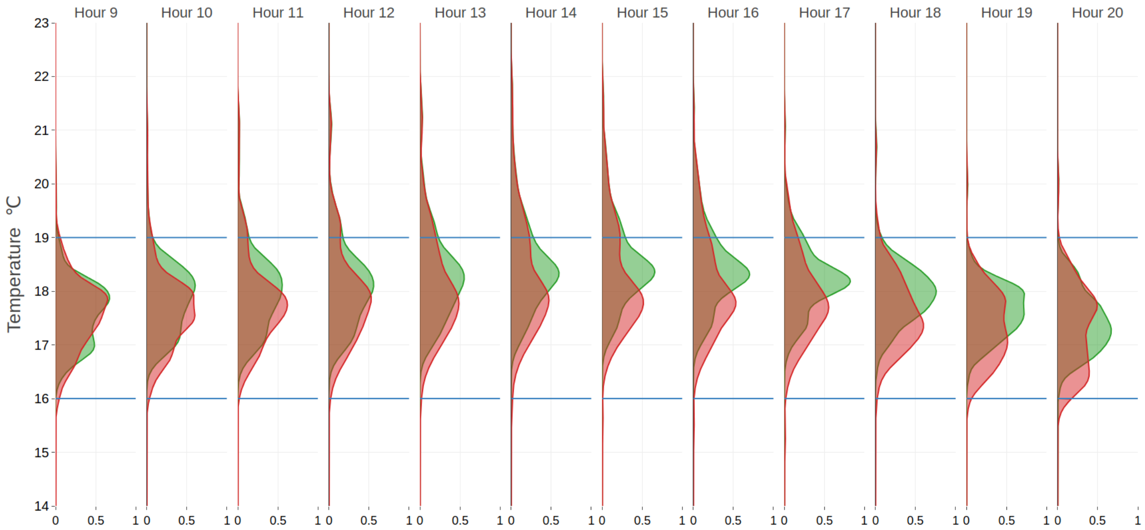

*Figure 1: Example of the temperature reached by the MS-FQI-1 (red) and PI-Const-15 control (green) in the working day of the winter testing period (*Results *paragraph). The areas show the empirical pdf of the Sales zone temperature, during the opening time, grouped by hour of the day.*

ergyPlus model, it generates the state observed by the agent by adding the past and forecast signals values, according to the Table 4. Then, the environment is in charge of restricting the set of action the agent can choose among. Furthermore, the environment receives the chosen action of the agent, it translates the action into a set-point and it sends back the set-point to the EnergyPlus model using the BCVTB interface. The simulation time-step of EnergyPlus is set to 5 minutes, while the delta-time between two consecutive RL time-step is fixed at 15 minutes. Differently from the other works of Ruelens et al. (2015) and Wei et al. (2017), the global set of actions available to the agent depends on the actual time of the day and on the actual season. The agent has a predefined set of action available on each season, that comprehends a only-ventilation action. In addition, the agent is allowed to choose a switch-off action before the opening time of the Sales zone, to achieve energy saving. The agent is in charge of choosing the action at each time-step and updating the Q-function approximation at the end of each episode. In the simulation analyzed, an episode is described as a day of activity of the supermarket, between the 8:30 a.m. and the closing time of that day.

Results

To demonstrate the performance of RL on retrofit applications some results are presented about simulations performed using the two modified supermarket models described before. The simulations cover two years, specifically from the 1^{st} January 2016 to the 30^{th} December 2017. The first year of each simulation is called *training period*, while the second year is called the *testing period*. The analysis of the simulations performed comprises only the testing period.

The baseline is defined using RBC control with constant seasonal set-point together with the PI control. The tuning of the baseline set-point follows the criterion of the stochastic constraint in Table 5. To accomplish the constraint requirements, the set-point of the RBC is fixed to $24°C$ for summer and $16.5°C$ for winter. The turn-on time of the HVAC system for the baseline is varied in order understand better the PI control behaviour with respect to the comfort and energy saving objectives. Therefore, the control named *PI-Const-30* turns on the HVAC at the 8:30 a.m. of each day, i.e. 30 minutes before the opening time of the Sales zone, *PI-Const-15* 15 minutes before and *PI-Const-0* turns on the plant at 9:00 a.m.

The RL agent using the Multi-Step Fitted-Q-Iteration algorithm is called MS-FQI-λ, depending on the λ parameter used in the reward function. In particular, λ is varied within the set $\{0.1, 1, 2\}$. The multi-Step parameter K is fixed to 10 and the discount factor $\gamma = 0.8$. The actions available to the MS-FQI are a discrete set of actions. Specifically, during winter the agent can choose from the set-point of $15°C$ to the set-point of $18°C$, with a step of $0.5°C$ (i.e. $15°C, 15.5°C, 16°C$ and so on). During summer, the MS-FQI is allowed to choose from the set-point of $24°C$ to $26°C$, with a step of $0.5°C$. The ϵ parameter of ϵ-greedy policy is linearly decreased from 1.0 to 0.02 during each RL time-step of the first 6 months of the training period, while is kept constant from that time-step on. The agent concentrates the greatest exploration during the training period while a little exploration is kept during the testing phase to continuously improve the policy.

To compare the simulations, two metrics are defined. The consumption metric is expressed in primary energy consumed during each season of the testing period, expressed in kWh/m^2. The maximum exceeding quantities mE used previously to define the

Table 6: Comparison of the results for climate zone B, testing period.

Control	$Cons$ winter	mE winter (%)			$Cons$ summer	mE summer (%)		
	(kWh/m^2)	$< 16°$	$< 15.5°$	$< 15°$	(kWh/m^2)	$> 25°$	$> 25.5°$	$> 26°$
PI-Const-30	24.99	0.91	0.00	0.00	40.54	7.45	2.03	0.36
PI-Const-15	**24.47**	**0.45**	**0.00**	**0.00**	**40.03**	**7.51**	**1.85**	**0.28**
PI-Const-0	25.28	0.56	0.01	0.00	41.21	15.81	14.04	12.39
MS-FQI-0.1	27.17	3.41	0.63	0.08	40.23	7.30	1.70	0.25
MS-FQI-1	20.92	2.36	0.09	0.00	39.12	15.33	5.59	1.32
MS-FQI-2	**19.51**	**7.71**	**0.85**	**0.21**	**38.92**	**16.60**	**6.31**	**1.61**

Table 7: Comparison of the results for climate zone E, testing period.

Control	$Cons$ winter	mE winter (%)			$Cons$ summer	mE summer (%)		
	(kWh/m^2)	$< 16°$	$< 15.5°$	$< 15°$	(kWh/m^2)	$> 25°$	$> 25.5°$	$> 26°$
PI-Const-30	34.89	18.49	7.77	2.37	35.17	11.76	4.76	1.80
PI-Const-15	**33.57**	**18.08**	**7.68**	**2.36**	**34.90**	**13.79**	**5.78**	**2.19**
PI-Const-0	32.61	18.28	8.11	3.01	35.41	34.88	27.25	21.63
MS-FQI-0.1	33.94	9.48	3.12	0.89	35.06	12.69	5.08	2.20
MS-FQI-1	31.56	14.69	6.60	2.50	34.94	14.72	5.42	2.17
MS-FQI-2	**30.38**	**17.40**	**7.96**	**3.01**	**34.38**	**16.85**	**5.81**	**2.14**

stochastic constraint are the comfort metric. As an example, refers to Figure 1: for each hour h, the area below $16°C$ is $E^h_{<16}$. The maximum over all the hours is the comfort measure $mE_{<16}$ (mE = maximum exceeding). For both the control systems, the max value is reached during 8 p.m. The MS-FQI-1 keeps a lower temperature than the PI-Const-15 but it remains in the comfort bounds, thus reducing the primary energy consumption. The area above $19°C$ is due to mid-season temperatures and because the HVAC is in heating mode during winter.

The results of the simulations are reported in Table 6 and Table 7 respectively for climate zones B and E. The results show that the performance of the PI-Const control systems vary with respect to the turn-on time-step. This demonstrates that the turn-on time is critical to fulfil the comfort constraints. The PI-Const-0 exceeds the stochastic constraint, while the other two baselines meet the thermal comfort requirements. The PI-Const-15 is chosen as best baseline because of better primary energy consumption.

The MS-FQI-λ agents satisfy the stochastic constraint. Note that during winter the mE values are greater for climate zone B while are smaller for climate zone E with respect to the three baselines. The MS-FQI-2 achieves a better energy saving for both the climate zones. Considering the overall testing period, it achieves an energy saving of 9.4% and 5.4% with respect to PI-Const-15, respectively for climate zone B and E. Overall, the savings obtained by the agent are higher for winter (20.3% zone B; 9.5% zone E), than for summer (2.8% zone B; 1.5% zone E).

Discussion

The overall results obtained for both climate zones are comparable with the primary energy saving obtained by Ruelens et al. (2015), even though their experiments used a very simplified thermal model. Furthermore, the RL agent is trained only for 12 months, differently from the previous work of Wei et al. (2017), where the training period consisted of 100 months.

The analysis performed shows that an RL control is a viable solution for retrofitting, specifically where design values are not sufficient anymore to guarantee thermal comfort requirements, because of degradation of HVAC system or change in building destination. In addition, reductions of the energy requests can have the same effect of over-sizing of the HVAC system, so it causes worsening of the control efficiency, for example after increasing thermal insulation of the envelope. In this case an adaptive algorithm like RL can reach greater savings with respect to an RBC control tuned for the original designated use and climate zone. As RL control learns by interaction with environment, it can achieve savings in every climate zone. To maximize energy saving and obtain near-optimal control, a re-tuning of the λ trade-off parameter should be taken into account when changing climate zone. An interesting improvement over the studied algorithm regards the reduction of the training period, which here consisted of 12 months. The reduction can be directed by an informed exploration policy, different from the ϵ-greedy used in the reported simulations, or by a pre-training phase in a simulated environment, as suggested by Wei et al. (2017). Model simulations do not account for noisy signals that can affect the performance of each control.An example is the weather forecast, which is assumed to be exact in the simulations performed while it is affected by error in common working conditions. More investigations are needed to assess the robustness of the RL control when receiving noisy observations and the adaptability with respect to varying climate zones and weather conditions. A first analysis of a retrofitting application of the RL framework on a real site showed promising results in terms of primary energy saving.

Conclusions

This work studied the performance of a RL algorithm with respect to state-of-the-art RBC control system in a realistic scenario, represented by two building models based on the Supermarket Reference Building provided by DOE. The overall results show that the the RL algorithm achieves 9.4% and 5.4% of primary energy saving respectively in climate zone B and E, while guaranteeing the same thermal comfort of RBC control. The same algorithm is applied as it is to different climate zones, without any tuning phase. This is an example of the scalability of the RL approach to different buildings. The research in ML offers opportunities to study and test new algorithms as building energy management systems. This study is a step forward to the application of one machine learning algorithm as a scalable solution for retrofitting.

Acknowledgments

We acknowledge that the present research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018-2022 (E11G 18000350001). The research leading to these results has been founded by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

References

ARPA (2018). Regional Agency for Environment Protection, Lombardia and Sicilia www.arpalomardia.it, www.arpa.sicilia.it. On-line resources.

ASHRAE (2017). Chapter 7: Fundamentals of Control. In *ASHRAE Handbook Fundamentals* (SI ed.). Atlanta, Ga. : ASHRAE.

Barrett, E. and S. Linder (2015). Autonomous HVAC Control, A Reinforcement Learning Approach. In *Machine Learning and Knowledge Discovery in Databases*, Cham, pp. 3–19. Springer International Publishing.

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba (2016). OpenAI Gym.

D.P.R. (1993). Decreto del Presidente della Repubblica, Italy. 26 agosto 1993, n. 412.

Ernst, D., P. Geurts, and L. Wehenkel (2005). Tree-based batch mode Reinforcement Learning. *Journal of Machine Learning Research 6*, 503–556.

Geurts, P., D. Ernst, and L. Wehenkel (2006). Extremely randomized trees. *Machine Learning 63*(1), 3–42.

MISE (2008). Decreto 11 marzo 2008, Ministero dello Sviluppo Economico Italy.

MISE (2015). Decreto 26 giugno 2015, Ministero dello Sviluppo Economico, Italy. Requisiti minimi degli edifici. [Minimum building requirements].

Page, J., D. Robinson, N. Morel, and J.-L. Scartezzini (2008). A generalised stochastic model for the simulation of occupant presence. *Energy and buildings 40*(2), 83–98.

Prez-Lombard, L., J. Ortiz, and C. Pout (2008, 01). A review on buildings energy consumption information. *Energy and Buildings 40*, 394–398.

Ruelens, F., S. Iacovella, B. Claessens, and R. Belmans (2015). Learning agent for a heat-pump thermostat with a set-back strategy using model-free Reinforcement Learning. *Energies 8*(8), 8300–8318.

Stefanutti, L. (2001). *Impianti di climatizzazione. Tipologie applicative [HVAC Systems. Applications].* Tecniche nuove.

Sturzenegger, D., D. Gyalistras, M. Morari, and R. S. Smith (2016). Model predictive climate control of a Swiss office building: implementation, results, and cost–benefit analysis. *IEEE Transactions on Control Systems Technology 24*(1), 1–12.

Sutton, R. S. and A. G. Barto (1998). *Introduction to Reinforcement Learning*, Volume 135. MIT press Cambridge.

Ente Nazionale Italiano di Unificazione (1995). *Aeraulic systems for wellness purposes. Generalities, classification and requirements. Rules for the offer request, offer, order and supply (UNI-10339).*

Ente Nazionale Italiano di Unificazione (2017). *Buildings energy performance (UNI-15232).*

Vázquez-Canteli, J. R. and Z. Nagy (2019). Reinforcement Learning for demand response: A review of algorithms and modeling techniques. *Applied Energy 235*, 1072–1089.

Watkins, C. (1989). *Learning from delayed rewards.* Ph. D. thesis, King's College, Cambridge, UK.

Wei, T., Y. Wang, and Q. Zhu (2017). Deep Reinforcement Learning for building HVAC control. In *Proceedings of the 54th Annual Design Automation Conference 2017.*

Wetter, M. (2011). Co-simulation of building energy and control systems with the Building Controls Virtual Test Bed. *Journal of Building Performance Simulation 4*(3), 185–203.

Zhang, Z., A. Chong, Y. Pan, C. Zhang, S. Lu, and K. P. Lam (2018). A Deep Reinforcement Learning approach to using whole building energy model for HVAC optimal control. In *2018 Building Performance Analysis Conference and SimBuild.*